The Dissertation Committee for Pengfei Zhang
certifies that this is the approved version of the following dissertation:

Optimization Models and Management Strategies for Service Operations

Committee:

Stephen M. Gilbert, Co-Supervisor

Douglas J. Morrice, Co-Supervisor

Diwakar Gupta

Jonathan F. Bard

Optimization Models and Management Strategies for Service Operations

by

Pengfei Zhang

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2021

To my parents.

Acknowledgments

While many students are lucky to have one great advisor, I am very fortunate to have two. This dissertation would not be possible without the steadfast supports and tireless efforts of my two advisors, Prof. Stephen M. Gilbert and Prof. Douglas J. Morrice. I am deeply grateful for their guidances and kindness when I have difficulties and confusion. We have also shared many joys of achieving goals, and this has made my Ph.D. journey enjoyable and memorable.

I would also like to thank the other two members of my committee, Prof. Diwakar Gupta and Prof. Jonathan F. Bard, for their valuable support of my research and study.

I have also benefited from many people's wisdom and knowledge that I have been able to learn from classes, friendships, conversations, or even from social connections. I sincerely express my special thanks to them all.

Optimization Models and Management Strategies for Service Operations

Publication No. _____

Pengfei Zhang, Ph.D.
The University of Texas at Austin, 2021

Supervisors: Stephen M. Gilbert
Douglas J. Morrice

In this dissertation, we investigate operational issues and strategies for three different service systems. In the first essay, we study how decentralized customer flows of the shared mobility service lead to the imbalance between the vehicle supply and customer demand, as well as intervention strategies for the network provider to improve the system. In the second essay, we study the appointment scheduling problem for integrated practice units using a two-stage integer stochastic programming model. In the third essay, we first study a data-free and distribution-free statistical characterization of random variables and then utilize it to design the order statistic uncertainty set for robust optimization that can be used to deal with uncertainty in service operations. We demonstrate our approach on the portfolio selection problem and the results demonstrate that our approach has superior performance relative to other uncertainty sets. In short, the studies in this dissertation address core issues in service operations, e.g., matching supply with demand and dealing with uncertainty. We provide both managerial insights and new optimization decision-making models that can be applied to improve service operations.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Fueled by service innovation and/or development of technologies, various service industries have undergone significant transformations. These transformations have created new challenges of managing the service operations, and this motivates us to study operational issues and strategies for service systems. We study three separate settings in which new operational challenges have been created as the result of new technologies. We develop both managerial insights and optimization decision-making models that can be applied to improve service performance. Our analyses address several core issues in operations analysis: in the first essay, we study matching supply with demand in a shared mobility network; in the second essay, we study scheduling optimization for integrated practice units; in the third essay, we study dealing with uncertainty with an application for portfolio optimization.

Essay 1: Managing Vehicle Flows in a Shared Mobility Network

Thanks to the development of internet and mobile technologies, the shared mobility service has emerged as a new and rapidly growing business model. In the first essay, we analyze the operations of the shared mobility service and develop strategies to improve the system. This research project was inspired by the observation that the BCycle bike sharing company used trucks to reposition bikes around the UT Austin campus. We realized that the decentralized customer flows in the bike sharing network could easily lead to the imbalance of bike supply and customer demand, and the company had to intervene to improve the bike circulations. We have interacted with firms from Austin and New York, and they confirmed that it is very challenging for such a decentralized

system to operate efficiently and that repositioning bikes is essential to their operations.

Intrigued by our discussions with BCycle's and Ofo's managers, we decided to investigate how the vehicle supply and customer demand in the shared mobility network are mismatched, and then develop strategies for the network provider to address the issue. We provide a new perspective on analyzing the sharing economy by investigating the implication of decentralized customer flows on the network operations. The most central element in our study is the critical location in the network whose total outbound demand constrains the total network flow. Interestingly, the notion of the critical location came up in our coffee conversation with Car2Go's North American Region Director, who mentioned that in their car sharing networks, there often exists a certain location where cars can get easily stranded because more customers go to the location than leaving there. The critical location also plays a pivotal role in our further analysis of the network equilibrium. We demonstrate a paradox which states that (under mild conditions) if any demand(s) to the critical location increases, then the total network flow will decrease, and excluding such a location could increase the total network flow. We also show that increasing the bottleneck location's total outward demand has a multiplier effect on the total network flow. Therefore, when vehicle repositioning is used to increase the network flow, the vehicles at the critical locations should be prioritized to reposition. As more vehicles are repositioned, new critical locations can emerge, and consequently, the company should also reposition the vehicles from new critical locations.

Essay 2: Appointment scheduling for integrated practice units

To improve delivery of care, many healthcare services are transitioning to value-based, patient-centered approaches, with improved coordination amongst providers. An Integrated Practice Unit (IPU) is a new approach to outpatient care in which a co-located multidisciplinary team of clinicians, technicians, and staff provide treatment in a single patient visit. While IPUs are structurally

organized around patients' needs with dedicated multidisciplinary teams, they present an operational challenge to coordinate activities among all providers. Effective coordination among all providers is needed to prevent delays and congestion, and thus scheduling becomes central to the efficiency of the IPU.

In this essay, we study the appointment scheduling problem to minimize a combination of closing time and total patient waiting time for the joint pain IPU at the Dell Medical School. We first present a new deterministic integer programming model for an extended open shop problem that can be used for clinic appointment scheduling for IPUs. We then discuss the advantages of the new model and introduce several valid inequalities to tighten the linear programming relaxation. To account for the stochastic element of the system, we further develop a two-stage integer stochastic programming model to determine the optimal appointment schedule for the IPU. The expected value solution is used to generate two different patient arrival templates, which are shown to be good candidates for assigning appointment times depending on whether the clinic closing time or the patient waiting time is the more important consideration. Sensitivity analysis confirms that the clinic statistics are stable for marginal changes in key resources.

Essay 3: Robust Optimization with Order Statistic Uncertainty Set with Application to Portfolio Selection

Dealing with uncertainty is a major challenge for making decisions in service systems, as well as in the general context of managing operations. Robust optimization has been a popular approach to address decision-making problems under uncertainty. However, when dealing with uncertainties, most existing robust optimization methods characterize random variables individually rather than *collectively*. We use the Probability Integral Transform to study a data-free and distribution-free statistical characterization of random variables: a set of i.i.d. random variables tend to be

scattered between extreme values rather than take extreme values.

To exploit the above idea for decision making under uncertainty, we have designed the *order statistic uncertainty set* for robust optimization. We match a set of random variables with different quantile levels so that different random variables can have different degrees of uncertainties. In this way, we can capture richer information from available data by utilizing the quantiles of random variables. We adopt the formulation of the assignment problem to develop a tractable formulation for the robust optimization model with the order statistic uncertainty set. The new order statistic uncertainty set provides a framework that incorporates the interval uncertainty set, the budget uncertainty set, and the demand uncertainty set as special cases. We report computational results on the portfolio selection problem with shortfall constraints to demonstrate that the order statistic uncertainty set has superior performance relative to other uncertainty sets.

Chapter 2

Managing Vehicle Flows in a Shared Mobility Network

2.1 Introduction

Shared mobility services broadly encompass services that allow users to access bikes, scooters, cars or other travel modes. With a variety of environmental and transportation-related benefits, the shared mobility services have been booming over the past decade. The National Association of City Transportation Officials (NACTO 2017) reported that bike-sharing trips in the U.S. have been growing steadily since 2010, with 35 million trips in 2017, 25% more than in 2016.

We study the local shared mobility service, where the service network is relatively small. In order for the network to make profits and achieve a high service level for customers, it is important for the vehicles to be available at the location where the service is requested. To ensure this, the network operator must strategically design the service region and deploy a proper amount of vehicles in the network. Once the vehicles begin to circulate in the network, the network flows will largely be determined the customers' decentralized movements, and the network operator must decide whether and how to intervene to increase the network flow.

A unique feature of the shared mobility system is its decentralization in the sense that the vehicle circulations entirely depend on the customer flows if the network operator does not intervene. The demand for movements between locations in the network are determined entirely by decentralized consumer flows. As the vehicles in the network are redistributed in this fashion, the vehicle supply of each location is determined by incoming customer flows from other locations. So when the number of vehicles at a location exceeds the customer demand there, the vehicles cannot all be

returned to the rest of the network. Unlike in the ridesharing service (e.g., Uber) where the vehicle can be relocated by the driver if there are excessive vehicle supplies at a location, the vehicle in the shared mobility system cannot move itself after each use. If the decentralized customer flows keep driving too many vehicles to the location with excessive vehicles, the vehicles will get stuck there. Because the customer demand from this location is not able to send out all vehicles to the rest of the network, it limits the vehicle supply to the rest of the network. As a result, some locations have excessive vehicles while other locations have unsatisfied demands, which affects the vehicle circulations in the network. Note that we will assume that the network operator deploys enough vehicles in the network (unless specified otherwise), i.e., any possible unsatisfied demands are not caused by inadequate vehicle supplies.

Our goal is to understand how the vehicle supply and customer demand are matched and/or mismatched without the network operator's intervention, which will then guide us to develop strategies for the network operator to address potential issues. For example, without the network operator's intervention, can the decentralized customer flows self-balance and keep vehicles circulating? If so, then we will refer to the self-balancing network flow as the equilibrium flow. For a given number of vehicles, would different initial placement of vehicles result in different equilibrium flows? Would supply and demand be perfectly matched in the equilibrium flow? If not, then what are the characteristics of the mismatch? What is the critical factor in the network that causes the mismatch between supply and demand, and how does the critical factor influence the total network flow? Is it worthwhile for the network operator to exert efforts (e.g., by repositioning vehicles) to increase the network flows, and what is the network operator's optimal strategy? Considering the increasing popularity of the shared mobility service, to understand these operational questions is vital to the shared mobility system.

We aim to address the above issues in this paper and our main contribution can be briefly

summarized as follows. First, we show that the decentralized customer flows can converge to the equilibrium flow, but without the platform's intervention, there can simultaneously exist unsatisfied customer demands and idle vehicle supplies. Second, we identify a critical location whose outbound demand constrains the total network flow and we show that increasing the critical location's outbound demand has a multiplier effect on the total network flow. We also demonstrate a paradox regarding the critical location, which states that if the demand(s) to the critical location increases, then the total network flow decreases. Finally, we develop intervention strategies for the network operator to increase the network flow. We show that the network operator should exclude certain locations from the service region even if there is no fixed cost for including the location. When vehicle reposition strategy is used to increase the total network flow, the network operator should prioritize repositioning vehicles from the critical location. In the following, we describe our contribution in detail.

We first study how the system evolves to the equilibrium flow without the network operator's intervention and then we analyze how the supply and demand are mismatched in the equilibrium flow. To do this, we develop a stylized model in which in each period, the network experiences the demands for travel between all pairs of locations in the network. In this study, we aim to study the impact of the spatial demand structure on the operations of the shared mobility network, so we abstract away the temporal variations in the demand structure and assume that the demands are time-invariant. In each period, the total flow out of each location is equal to either the supply of vehicles at the location at the beginning of the period or the outbound demand, whichever is smaller. The availability of vehicles at the beginning of the period is governed by balance of flow constraints. When the total outbound demands at a location exceeds the availability of vehicles, we assume that the vehicles are allocated in proportion to the demand volume in different directions. The network operator initially deploys a certain amount of vehicles in the network

and the vehicles will be redistributed by customer flows afterwards. We find that no matter how the network operator deploys the vehicles initially, the network flow eventually converges to the same equilibrium flow. We develop a linear model to characterize the equilibrium flow that will be achieved without the intervention of the network operator. We find that without the network operator's intervention, the equilibrium flow is not perfect because some locations can have idle vehicles while other locations have unsatisfied demands.

We then identify a critical location in the network that constrains the total equilibrium flow in the network as a result of a large imbalance between its inbound and outbound demands. In the equilibrium flow, there is usually a critical location whose total outbound flow reaches its upper bound, i.e., its total outbound demand. When the number of vehicles in the network exceeds the amount of the equilibrium flow, the vehicles at the critical location start to accumulate and all the vehicles in excess of the equilibrium flow ultimately reach and remain at the critical location. Eventually, the outbound flows of all other locations will not be able to get beyond that in the equilibrium flow.

Next, we demonstrate a paradox which states that if any demand(s) to the critical location increases, then the total network flow will decrease. The paradox occurs because if more vehicles are driven to the critical location, more vehicles will be stranded there and the number of vehicles in the rest of the network will decrease. This paradox implies that it may be possible to increase the total equilibrium flow in the network by eliminating a (critical) location. The paradox is also useful for evaluating the effect of how changes in the demands into the critical location can affect the total equilibrium flow in the network. We show that increasing the critical location's total outbound demand has a multiplier effect on the total network flow, i.e., the total network flow increases by at least twice as much as the increase of the critical location's total outbound demand. In contrast, the (local) change of any other location's total outbound demand does not affect the

total network flow at all.

Guided by the above equilibrium analysis of the network flows, we develop intervention strategies for operating the network. From the above, we know that the critical location's limited outbound demand not only constrains its own flow but also restricts the equilibrium flow throughout the entire network. This presents an opportunity for the network operator to intervene. We study two possible strategies for the network operator: eliminating the critical location or enhancing the outbound demand of the critical location by repositioning vehicles. We show that even if there is no fixed cost for including a location in the network, eliminating the critical location may increase the amount of the equilibrium flow in the network. Because the critical location's outbound demand is the key constraining factor for the total network flow, the network operator can enhance the critical location's outbound flows by repositioning vehicles from it. With a hub-and-spoke network based on the real data from the BCycle bike sharing company, we study the insights for the network operator's reposition strategy. We find that prioritizing repositioning vehicles from the critical location is optimal. As more vehicles are repositioned from the critical location, new critical locations can emerge, and consequently, the network operator should also reposition the vehicles from the new critical locations.

The remainder of our paper is organized as follows. In section 2.2, we provide a brief literature review. In section 2.3, we present the network model for the shared mobility system and study the system evolution. In section 2.4, we analyze the system equilibrium flow. In section 2.5, we study two intervention strategies for the network operator to improve the network.

2.2 Literature Review

Our work is related to the rapidly growing literature that study the operations of the shared mobility service. One stream of related literature aims to identify locations that have significant imbalance

between the inbound and outbound vehicle flows. For example, Hong et al. (2015) and Jin et al. (2016) proposed detection algorithms to identify the black-hole location, which is defined as the location whose ratio of the inbound flow to the outbound flow exceeds a threshold. Their characterizations are based on an ad hoc measure, and their analysis is limited to be at the local level because each location is examined as an individual component isolated from the network. In contrast, our analysis is situated at the systemic level and it captures the network interactions among different locations. The black-hole location identified based on their measure is fundamentally different from the critical location identified using our analytical model.

Our study analyzes the network operator's strategies for both long-term planning and short-term decision problems. We discuss related literature for each of them. Over the long term, the network operator must decide on the service region selection and/or fleet size. George and Xia (2011) used a queueing network model to determine the optimal fleet size for a given network of stations. They first derived the relation between the vehicle availability and the fleet size, based on which they optimized the fleet size. In some other studies, the service region selection and the fleet sizing are jointly optimized. For example, Lu et al. (2018) considered a strategic planning problem for car-sharing systems with a two-stage stochastic integer programming model, where the first stage determines the vehicle fleet allocation to different service zones and the second stage models the vehicle movements. They applied their approach to a real-world problem with a rolling-horizon framework and showed that their model leads to a substantial improvement of profitability and quality of service over an intuitive benchmark policy. He et al. (2017) studied both the service region planning and fleet sizing for the electric vehicle sharing system and developed a distributionally robust optimization model to incorporate the uncertainty with regard to customers' travel patterns. When optimizing the service region, these studies always assume fixed costs to serve different regions. However, in our study, we show that even if there are no fixed costs to

serve each region, certain regions still need to be excluded from service in order to maximize the total network flow.

Once the service region and the fleet size are determined, the operational decision problems come up. Banerjee et al. (2016) studied using dynamic pricing to control the shared vehicle system for various objectives. The focus of their study is to design efficient algorithms and develop approximation guarantees for the problems. The study in Shu et al. (2013) found that the dynamically repositioning vehicles in the network with the time-varying demand affects both the service level and the utilization of vehicles, and the effectiveness of the vehicle reposition depends on the demand usage patterns and the number of vehicles in the network. Also related is a group of papers that studied the route optimization for vehicle reposition, e.g., Cruz et al. (2017) and Elhenawy and Rakha (2017). These papers usually assume that the origin locations and destination locations of repositioned vehicles are given, and the problems in these papers are essentially classical network problems being applied in the shared mobility system; for example, the problem in Cruz et al. (2017) is a minimum-cost flow problem.

The first unique analysis in this paper is to study how the decentralized customer flow is able to reach the equilibrium flow, and to show how the supply and demand are mismatched in the equilibrium flow. Another part of our contribution is that we identify a critical location in the network and demonstrate the multiplier effect of its outbound demand on the total network flow, as well as a paradox related to the critical location. Finally, we provide insights for the service region design and develop strategies for the network operator to reposition vehicles.

2.3 The Network Model

We focus our study on local shared mobility networks. Assume there are $n \geq 2$ locations, indexed $1, \cdots, n$, for which there is an infinite horizon of discrete-time periods of time invariant demands.

Consider an infinite-horizon discrete-time problem where the demand is deterministic and time-invariant. At the beginning of period $t = 1$, the network operator distributes a specific number of vehicles among the locations in the network. We assume that the total number of vehicles is large enough that this number does not constrain the total flow in the network. It will become clear later how many vehicles are necessary for this.

At the beginning of each period, there are Q_{ij} potential customers seeking to go from location i to location j; let $Q_{ii} = 0, \forall i = 1, \cdots, n$. We call $\boldsymbol{Q}$ the network's demand pattern, where the ij-th entry of matrix $\boldsymbol{Q}$ is Q_{ij}. We assume that the network is complete (unless specified otherwise), i.e., there is a directed arc from each location to all other locations and the demand on every arc is strictly positive.

For many local shared mobility services, the price for the trip is usually set to be the same if the trip duration is within a time threshold. For example, in the BCycle bike-sharing system, the price for trips within 60 minutes is the same, and extra cost incurs for each additional 30 minutes. Because customers of local shared mobility services primarily use vehicles for short-term use, most trip durations typically fall within the time threshold set by the network operator. The data for BCycle trips in Austin shows that about 90% of bike trips last no more than the threshold (60 minutes), and thus most trips yield the same revenue. Therefore, for simplicity, we assume that the revenue for the trip on every arc is the same. As a result, in order to analyze the network revenue, we can simply study the total network flow.

In local shared mobility networks, the price for the trip is often set to be the same if the trip duration is within a time threshold. Because the customers primarily use vehicles for short-term use, most trip durations typically fall within the time threshold set by the network operator. For example, in the BCycle bike-sharing system, the price for trips within 60 minutes is the same, and extra cost incurs for each additional 30 minutes. The data for BCycle trips in Austin shows that

about 90% of bike trips last no more than the threshold (60 minutes), and thus most trips yield the same revenue. Therefore, for simplicity, we assume that the revenue for the trip on every arc is the same. As a result, in order to analyze the network revenue, we can simply study the total network flow.

We assume that each trip takes exactly one period and the vehicle used in the current period will be available to use again from the destination at the beginning of the next period. In every period, the outbound vehicles leave each location before the inbound vehicles from other locations arrive to the location. In each period, the total flow out of a location is equal to the minimum of the number of vehicles at that location, and the total demand from that location to all other locations. If the number of vehicles at the location at the start of the period is not enough to satisfy all of the outbound demand, then we assume that the supply of vehicles is allocated among the outbound arcs in proportion to their volumes of demand. Specifically, we assume that the fraction of the vehicle supply at location i that is allocated to arc (i, j) is $\alpha_{ij} = \frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}}$. This assumption is consistent with a random selection of departures from location i to which to assign vehicles.

We use the following notation to describe the system state in each period.

$I_i(t)$: Ending inventory of vehicles at location i at the end of period t

$Y_{ij}(t)$: Number of vehicles leaving from location i to location j in the beginning of period t

$X_j(t)$: Total number of vehicles leaving from location i to other locations in the beginning of period t; we have $X_j(t) = \sum_k Y_{jk}(t)$.

Suppose the initial system state is $I_1(0), \cdots, I_n(0)$. Let $X_j(0) = Y_{ij}(0) = 0, \forall i, j = 1, \cdots, n$. For $t \geq 1$, the network evolves according to the following Decentralized Vehicle Movement (DVM) protocol.

(a) **Demand fulfillment.** At the beginning of every period t, the total flow out of the

location i is equal to the minimum of the number of vehicles at location i, and the total outbound demands from location i.

$$X_j(t) = \min\left\{\sum_{k=1}^{n} Q_{jk},\ I_j(t-1)\right\},\ \forall j = 1,\cdots,n. \tag{2.1a}$$

(b) **Proportional vehicle movement.** At time period t, if the supply of vehicles at node i is less than the total demand for travel from node i to all other nodes, then the available supply of vehicles is allocated to the arc (i,j) in the proportion α_{ij}.

$$Y_{ij}(t) = X_i(t)\cdot\alpha_{ij},\ \forall i,j = 1,\cdots,n,\ \forall t. \tag{2.2a}$$

This assumption is similar to what has been used in other vehicle sharing models, including Shu et al. (2013) and He et al. (2017).

(c) **Inventory balance condition.** Each location maintains and carries forward in inventory any unused vehicles in the period.

$$I_j(t) = I_j(t-1) - X_j(t) + \sum_{i=1}^{n} Y_{ij}(t). \tag{2.3a}$$

A system state $\boldsymbol{x}$ is an equilibrium flow if once the system is in state $\boldsymbol{x}$, it will remain in state $\boldsymbol{x}$. In general, not all linear dynamic systems can reach a unique equilibrium flow. However, as we will prove later, the linear system (2.1) - (2.3) always evolves into a unique equilibrium flow for a given number of vehicles. We first study the following problem that computes the maximum possible network flow that self-balances.

2.3.1 The Equilibrium Flow

We first study the following linear model that will be used to characterize the network equilibrium flow. This linear model is also used in Shu et al. (2013) to study the fleet sizing in the equilibrium state of the bicycle sharing network.

$$\max_{x} \sum_{j} x_j \tag{2.4a}$$

$$s.t.\ x_j = \sum_{i=1}^{n} x_i \cdot \alpha_{ij}, \forall j = 1, \cdots, n \tag{2.4b}$$

$$0 \leq x_j \leq \sum_{k=1}^{n} Q_{jk}, \forall j = 1, \cdots, n, \tag{2.4c}$$

Proposition 1. *If the network is strongly connected, then Problem* (2.4) *has a unique optimal solution* $\boldsymbol{x}^*$.

In fact, any feasible solution of the above flow maximizing problem is a multiple of $\boldsymbol{x}^*$. Therefore, $\boldsymbol{x}^*$ can be seen as the characteristic solution with respect to the demand pattern $\boldsymbol{Q}$. As we will see later, the solution $\boldsymbol{x}^*$ plays a central role in characterizing the system evolution.

We can easily see another important property of the optimal solution to Problem (2.4): there is at least one location j that satisfies $x_j^* = \sum_{k=1}^{n} Q_{jk}$ because otherwise, we can scale up the solution to increase the objective function. We refer to this location as the *critical* location, and later we will show that (1) it drains all excessive vehicles during the network evolution; (2) its outbound demand constrains the equilibrium flow; (3) it plays a key role in the network operator's intervention strategy. In the following, we assume there is only one critical location in the network (unless specified otherwise), but the results can be easily generalized to cases with more than one critical locations.

2.3.2 System Convergence and Globally Stable Equilibrium

Suppose that at the beginning of time $t = 1$, the network operator deploys $\eta \cdot \sum_{j=1}^{n} x_j^*$ number of vehicles in the network, where $\eta > 0$. The system evolves according to the DVM protocol. For the time period t, define the following ratio for location j: $r_j(t) = \frac{X_j(t)}{x_j^*}$, $\forall j = 1, \cdots, n$. Denote $\hat{r}(t) = \min\left\{r_1(t), r_2(t), \cdots, r_n(t)\right\}$ and $\Delta\hat{r}(t) = \min\{\eta, 1\} - \hat{r}(t)$.

Proposition 2. *Suppose the network is complete and has at least 3 locations. The following results hold.*

1. *$\hat{r}(t)$ (weakly) monotonically increases in time t with $\lim_{t\to\infty} \hat{r}(t) = \min\{\eta, 1\}$.*

2. **Bound for the convergence rate:** $\min\{\eta, 1\} - \hat{r}(t+2) \leq C_r \cdot \left[\min\{\eta, 1\} - \hat{r}(t)\right]$, *where*

$$C_r = 1 - \min_{j_1, j_3 \in \{1, \cdots, n\}} \left\{ \frac{x_{j_1}^*}{x_{j_3}^*} \cdot \sum_{\substack{j_2 = 1 \\ j_2 \neq j_1, j_3}}^{n} \alpha_{j_1, j_2} \cdot \alpha_{j_2, j_3} \right\} < 1.$$

3. **Globally Stable Equilibrium**. *No matter how vehicles are placed initially, we have $\lim_{t\to\infty} X_j(t) = \min\{\eta, 1\} \cdot x_j^*, \forall j = 1, \cdots, n$.*

4. **The Draining Effect**. *If $\eta > 1$, then all the vehicles in excess of $\sum_{j=1}^{n} x_j^*$ will stay idle at the critical location as time $t \to \infty$.*

According to the third result, the system state always converges to the same globally stable equilibrium regardless of how vehicles are initially placed in the network; for example, the network operator can initially put all vehicles at any single location or spread the vehicles in the entire network. In order to achieve the maximum possible equilibrium flow $\boldsymbol{x}^*$, η should be at least 1, e.g., the network should have at least $\sum_{j=1}^{n} x_j^*$ number of vehicles. Moreover, the last result shows that the network operator does not need to deploy more than $\sum_{j=1}^{n} x_j^*$ number of vehicles because

the vehicles in excess of $\sum_{j=1}^{n} x_j^*$ would be drained to the critical location. So from now on, we assume the network operator always deploys $\sum_{j=1}^{n} x_j^*$ number of vehicles in the network (unless stated otherwise). Therefore, the maximum possible equilibrium flow $\boldsymbol{x}^*$ will simply be referred to as the equilibrium flow.

At time t, if a location j has a small (large) ratio $r_j(t)$, then in the next time period $t+1$, the location j tends to receive relatively more (less, respectively) flow from other locations, and thus its ratio $r_j(t+1)$ may go up (down, respectively). As a result, each location j's ratio $r_j(t)$ and outbound flow $X_j(t)$ will oscillate over time. Because $\min\{\eta,1\} - \hat{r}(t+2) \leq C_r \cdot \big[\min\{\eta,1\} - \hat{r}(t)\big]$ holds, such an oscillation is generally a damped one because it decays with time. The ratio $r_j(t)$ can be viewed as the "distance" between $X_j(t)$ and x_j^* at time t. The minimum ratio $\hat{r}(t) = \min\big\{r_j(t), \forall j = 1, \cdots, n\big\}$ measures how far the system's state is from the equilibrium flow. The value C_r provides a bound for the decay rate.

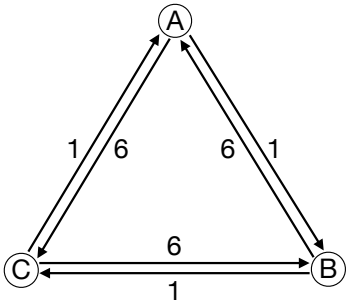


Figure 2.1: A network with 3 locations: A, B and C.

We illustrate the network flow evolution with the network in Figure 2.1. There are 3 locations and the demand for each arc is shown in the figure. The equilibrium flow $\boldsymbol{x}^* = (x_A^*, x_B^*, x_C^*) = (7,7,7)$. Suppose we initially place 4 units of vehicles at location A. Figure 2.2 shows how $X_A(t)$, $X_B(t)$ and $X_C(t)$ evolve over time. The state of each location exhibits a pattern of damped oscillation, which eventually delays to its limit. We can also see that from time $t = 2$, the location

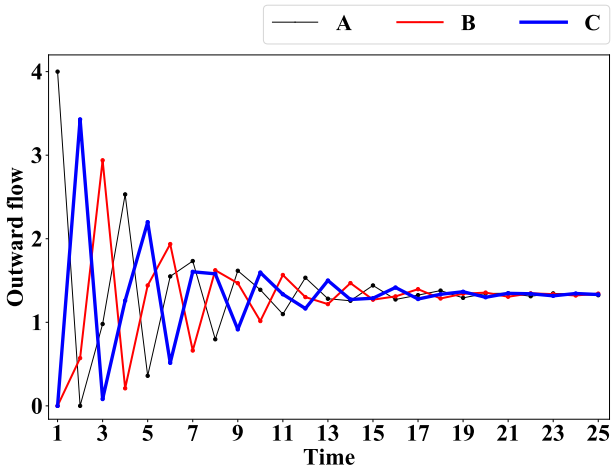


Figure 2.2: The system evolution exhibits a pattern of damped oscillation around the scaled equilibrium flow and eventually decays to the scaled equilibrium flow.

that achieves the minimum ratio $\hat{r}(t)$ shifts as follows: $A \to C \to B \to A \to C \to B \to A \to \cdots$.

2.4 Equilibrium Analysis

Since the system state always converges to the globally stable equilibrium, the total network flow in the long term is determined by the equilibrium flow. In this section, we analyze the equilibrium flow, which will provide useful guides for the network operator's intervention strategy.

2.4.1 Sensitivity Analysis

To evaluate the influences of different locations on the total network flow, we study how the network equilibrium flow is affected when the outbound demands of different locations change. In this Section 2.4.1, we assume there is only one critical location in the network.

We first study how the network equilibrium flow is affected by the critical location's outbound demands. Assume an infinitesimal quantity $\tau > 0$ and that the vector $\boldsymbol{\epsilon}_j = (\varepsilon_{j1}, \varepsilon_{j2}, \cdots, \varepsilon_{jn})$ satisfies $\varepsilon_{jj} = 0$, $\sum_{l=1}^{n} \varepsilon_{jl} = 1$ and $\varepsilon_{jk} \geq 0, \forall k \neq j$. We simultaneously increase the location j's outbound

demands as follows: its outbound demand Q_{jk} is increased to be $Q_{jk} + \tau \cdot \varepsilon_{jk}$, for all $k \neq j$. Note that the outbound demands of any other location l $(l \neq j)$ stay the same. Suppose the optimal objective value of problem (2.4) is changed by $\sigma(\tau, \epsilon_j)$ after the outbound demands of the critical location j are perturbed by infinitesimal amounts. Because the change of location j's total outbound demand is $\tau \cdot (\sum_{l=1}^{n} \varepsilon_{jl}) = \tau$, we can define the sensitivity coefficient of location j with respect to ϵ_j as $\lambda_j(\epsilon_j) = \lim_{\tau \to 0} \frac{\sigma(\tau, \epsilon_j)}{\tau}$.

Proposition 3 (Multiplier Effect)**.** *For any complete network, suppose it has only one critical location j, then for any ϵ_j that satisfies $\varepsilon_{jj} = 0$, $\sum_{l=1}^{n} \varepsilon_{jl} = 1$ and $\varepsilon_{jk} \geq 0, \forall k \neq j$, we have* $1 + \frac{1}{\max_k \alpha_{kj}} \leq \lambda_j(\epsilon_j) \leq 1 + \frac{1}{\min_k \alpha_{kj}}$.

The lower bound and upper bound for the critical location j become equal to each other if $\alpha_{ij} = \alpha_{kj}, \forall i \neq j, k \neq j$ hold. In the above sensitivity analysis, the perturbation of the outbound demands from the critical location j can be on one or more outbound arc(s) from the critical location j. Note that $1 + \frac{1}{\max_k \alpha_{kj}} \geq 2$, so the above result shows the multiplier effect of the critical location, i.e., the total network flow increases by at least twice as much as the increase of the critical location's outbound demand.

The lower bound for $\lambda_j(\epsilon_j)$ can become arbitrarily large (for networks with at least 3 locations). We illustrate this with the following example.

Example 1. *Consider the network in Figure 2.3 with locations A, B and C. The demand on each of the 6 arcs is shown in the figure. Suppose $\theta > 1$ and M is non-negative. The location A is the critical location and the total equilibrium flow is $2 \cdot (2 + M)$. Assume τ is an infinitesimal quantity and satisfies $0 < \tau < 2 \cdot \theta - 2$. Suppose each of the two outbound demands from location A changes from 1 to $1 + \frac{\tau}{2}$, then the maximum equilibrium flow becomes $(2 + \tau) \cdot (2 + M)$. In this case, location A's sensitivity coefficient $\lambda_A(\frac{1}{2}, \frac{1}{2})$ is equal to $\frac{\tau \cdot (2+M)}{\tau} = 2 + M$. We can see that $\lambda_A(\frac{1}{2}, \frac{1}{2})$ becomes*

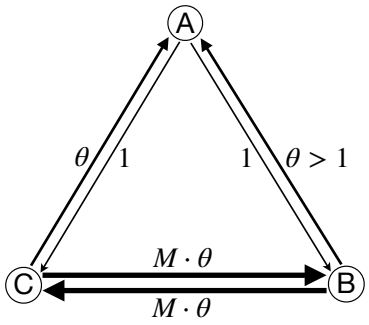


Figure 2.3: The sensitivity coefficient of the critical location A is equal to $2+M$.

2 if $M = 0$ *and* $\lambda_A(\frac{1}{2}, \frac{1}{2})$ *approaches infinity as M increases to infinity. Note that the lower bound and upper bound for the sensitivity coefficient of the outbound demands of the location A are both equal to* $2+M$. □

The multiplier effect is not an exclusive property of the critical location. When we increase demands between non-critical locations, there can also exist the multiplier effect and the corresponding sensitivity coefficient can also be arbitrarily large. We demonstrate it with an example. We first define the sensitivity coefficient of the demand change from location j to location k. Suppose the demand from location j to location k increases by $\tau > 0$, i.e., the demand from j to k becomes $Q_{jk} + \tau$, and consequently, the optimal objective value of problem (2.4) is changed by $\sigma_{j\to k}(\tau)$. Then we define the sensitivity coefficient from location j to location k as $\lambda_{j\to k} = \lim_{\tau\to 0} \frac{\sigma_{j\to k}(\tau)}{\tau}$.

Example 2. *In order to develop the intuition for the complete network, we first study the Z-shaped network in Figure 2.4 with locations A, B, C and D. The demand on each of the 6 arcs is shown in the figure. Suppose* $0 < \varepsilon_1 < \varepsilon_2 \approx 0$ *and* $M > 1$. *The equilibrium flow is as follows:*

$$\textit{The flow from A to B (or from B to A)} : (1+\varepsilon_1)\cdot(1-\varepsilon_2)\cdot M$$

$$\textit{The flow from B to C (or from C to B)} : (1+\varepsilon_1)\cdot(1-\varepsilon_2)$$

$$\textit{The flow from C to D (or from D to C)} : 1-\varepsilon_2$$

In the equilibrium, the location D is the critical location, and the total equilibrium flow is $(2-2\cdot\varepsilon_2)\cdot(2+\varepsilon_1+M+M\cdot\varepsilon_1)$. *By taking the partial derivative of the total equilibrium flow with respect to* ε_1, *we can get* $\lambda_{C\to B}=(2-2\cdot\varepsilon_2)\cdot(1+M)\approx 2+2\cdot M>4$, *which approaches infinity as M approaches infinity.*

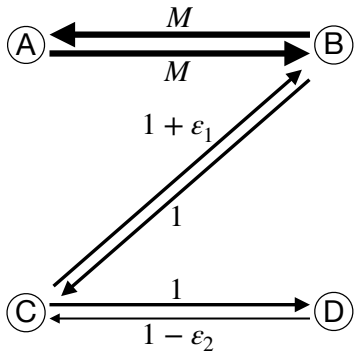


Figure 2.4: The location D is the only critical location; $\lambda_{C\to B}=(2-2\cdot\varepsilon_2)\cdot(1+M)\approx 2+2\cdot M$ increases as M grows.

We then present the numerical result of a complete network. For it, we add 6 arcs (between A and C, between A and D, between B and D) in the Z-shaped network in Figure 2.4. The demands on the 6 added arcs (not shown in Figure 2.4) are equal to 0.001; as we will see shortly, the 6 added arcs only have a very small impact on $\lambda_{C\to B}$ *because the demands on them are very small. We set* $M=100$, $\varepsilon_2=0.01$. *We consider two cases:* $\varepsilon_1=0.001$ *and* $\varepsilon_1=0.002$. *Our calculation shows that the location D remains to be the only critical location for both of the two cases. When* ε_1 *increases from 0.001 to 0.002, the total equilibrium flow increases from 202.1730 to 202.3720. The ratio* $\frac{202.3720-202.1730}{0.002-0.001}=199$ *measures the sensitivity of the total equilibrium flow with respect to the demand change from C to B. Note that the ratio 199 is approximately equal to* $\lambda_{C\to B}=(2-2\cdot\varepsilon_2)\cdot(1+M)=1.98\times 101=199.98$ *for the Z-shaped network in Figure 2.4.* □

In the above network in Figure 2.4, when the demand from C to B increases, more proportion of vehicles from C will go to the location B. As a result of the flow balance condition between B

and C, the flow from B to C will also increase. As a result of the proportionality condition of the location B, the flow from B to A will also increase and the increment of the flow from B to A is much larger than the increment in the demand from C to B.

Although the demands between non-critical locations can have a multiplier effect on the total equilibrium flow, they always have less influence than the demand from the critical location to non-critical locations. To illustrate it, we next compare the demand changes for the case where the outbound demand from the critical location increases and the case where the outbound demand from the non-critical location increases.

Proposition 4. *For any complete network with at least 3 locations, suppose location j is the only critical location and location i and k are two different non-critical locations, then we have* $\lambda_{j\to k} - \lambda_{i\to k} \geq 1 - \frac{x_i^*}{\sum_m Q_{im}} > 0$.

The above result shows that the demand increase from the critical location to any non-critical destination location is more beneficial than from any non-critical location to the same non-critical destination location. As we will prove later (in Proposition 7), when we add an infinitesimal amount of demand from the critical location j to k, the location j remains to be the critical locationn; this means that in the new equilibrium, the added demand will all get served and the flow from the location j will increase by the same amount. In contrast, if we add some demand from the non-critical location i to k, then only around $\frac{x_i^*}{\sum_m Q_{im}}$ of the added demand will get served. This leads to the difference between the above two sensitivity coefficients.

An important implication of the above result is that if the network operator wants to increase the equilibrium flow by repositioning vehicles, then the vehicles at the critical location should be prioritized to reposition. We will study more vehicle reposition strategies later in Section 2.5.2.

2.4.2 A Paradox

In the shared mobility system, the demand and supply coincide, i.e., the flow of satisfied demands in the current period is the flow of vehicle supply of the next period. When the demand on an arc increases, it could potentially makes the vehicle redistribution less efficient. As a result, the network's total flow may decrease if the demand on particular arc(s) increases. We provide such a paradox in this section.

Assume the set of critical locations as $J_S(\boldsymbol{Q})$. We add demands to the demand pattern $\boldsymbol{Q}$. Assume a nonzero $n \times n$ demand increment matrix Δ that satisfies the following condition: $\delta_{ii} = 0$, $\delta_{ij} \geq 0$ if $i = 1, \cdots, n, j \in J_S(\boldsymbol{Q})$, and all the remaining entries are zeros. Let S_1 be the set $\{j : \sum_i \delta_{ij} > 0, j = 1, \cdots, n\}$; we have $S_1 \subseteq J_S(\boldsymbol{Q})$. For any location in the set S_1, at least one of its inbound demand increases. Let $\boldsymbol{Q} + \Delta$ denote the new demand pattern where the demand from location j to k is $Q_{jk} + \delta_{jk}$. Denote the equilibrium flow for the demand pattern $\boldsymbol{Q} + \Delta$ as $\widetilde{\boldsymbol{x}}^*$.

Proposition 5 (A Paradox)**.** *Suppose the network is complete.*

(1) If the set $A_1 = \{j : \sum_i \delta_{ij} > 0, \sum_i \delta_{ji} > 0\}$ *is empty, then we have* $\widetilde{x}_j^* \leq x_j^*, \forall j = 1, \cdots, n$.

(2) Further if the set $A_2 = \{j : \sum_i \delta_{ij} = \sum_i \delta_{ji} = 0\}$ *is non-empty, then we have* $\widetilde{x}_j^* \leq x_j^*, \forall j \in S_1$*;* $\widetilde{x}_j^* < x_j^*, \forall j \in \{1, \cdots, n\} \setminus S_1$.

Compared with $\boldsymbol{Q}$, the demand pattern $\boldsymbol{Q} + \Delta$ has at least the same amount of demand on every arc and strictly more demand on at least one arc whose destination is a critical location. The first result in Proposition 5 shows that if the demands from any location(s) to any critical location increase and no location's inbound demands and outbound demands both increase, then in the new equilibrium flow $\widetilde{\boldsymbol{x}}^*$, no location's outbound flow (or inbound flow) will increase. The intuition is explained in the following. In the equilibrium flow of $\boldsymbol{Q}$, all outbound demands of the critical locations in the set S_1 are already satisfied. In the case of $\boldsymbol{Q} + \Delta$, the critical locations in the set

S_1 have more inbound demand, but they can only return as many vehicles to the network as in the case of $\boldsymbol{Q}$. Therefore, the critical locations in the set S_1 receive more vehicles but do not return more vehicles to the network. As a result, if a location sends more vehicles to a critical location in S_1, the extra vehicles will not be returned to the network once they reach the critical location. The locations in $\{1,\cdots,n\} \setminus S_1$ will receive the same or strictly less proportion of flows from any locations; so the inbound flow to any location in $\{1,\cdots,n\} \setminus S_1$ will not increase. Because the outbound flows of the critical locations in S_1 do not increase and the inbound flows of the location in $\{1,\cdots,n\} \setminus S_1$ do not increase, the entire network's total flow does not increase.

The second result in Proposition 5 shows that if there exists a location whose inbound demand and outbound demand both do not increase, then the outbound flow of any location in $\{1,\cdots,n\} \setminus S_1$ will strictly decrease. The intuition is explained in the following. Assume there is a location l_1, and both its inbound demand and outbound demand do not increase. Then the location l_1's inbound flow will decrease because the location l_1 receives less flow from the locations that send more flows to the locations in S_1. Therefore, the inbound flow of the location l_1 will strictly decrease, and it will send strictly less flow to other locations. For the locations in $\{1,\cdots,n\} \setminus S_1$, they receive no more flows from any locations in the network, and strictly less flow from the location l_1. Therefore, the inbound flow of any location in $\{1,\cdots,n\} \setminus S_1$ will strictly decrease, and so does the outbound flow.

From the perspective of the draining effect, if the demand into the critical location increases, then the draining effect is enhanced. In this sense, any incremental demand into any critical location can be viewed as the "bad" demand, and it makes the balancing of supply and demand less efficient.

If there is only one critical location in the network, Proposition 5 becomes the following simpler result. Note that we still assume that we only increase demands to the critical location.

Corollary 1. *Suppose the network is complete and the location j is the only critical location. If there exists at least one location m $(m \neq j)$ such that $\delta_{mj} = 0$, then we have $\widetilde{x}_j^* \leq x_j^*$ and $\widetilde{x}_k^* < x_k^*, \forall k \neq j$.*

2.4.2.1 Further Discussions on the Paradox

We demonstrate the paradox with a simple network in Figure 2.5, which consists of a hub s and two spokes 1 and 2. The locations 1 and 2 are not directly connected. We fix $Q_{1s} = 3$ and $Q_{2s} = 1$. Then we calculate the total network flow for different values of Q_{s1} and Q_{s2}, each ranging from 0 to 8. The left panel of Figure 2.6 shows the grayscale plot of the network revenue for different values of Q_{s1} and Q_{s2}. The darker points correspond to the cases with larger total network flows.

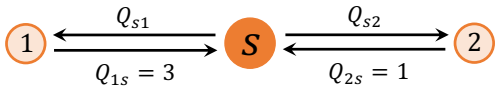


Figure 2.5: A network with 3 locations.

The maximum possible total network flow (equal to 8) will be achieved if $Q_{s1} = 3 \cdot Q_{s2}$ and $Q_{s2} \geq 1$ hold; in any other case, the network revenue is less than 8. We divide the two-dimensional space for Q_{s1} and Q_{s2} into three regimes as shown in the right panel of Figure 2.6. In the normal "sweet spot" regime, the revenue increases if either Q_{s1} or Q_{s2} increases. In the upper regime, the location 2 is the critical location; in this case, if the demand to the critical location (i.e., Q_{s2}) increases (keeping Q_{s1} unchanged), then the network revenue will decrease. In the lower right regime, location 1 is the critical location, and any increment of Q_{s1} decreases the network revenue. Note that when we add small amounts of demands between the location 1 and location 2 to make it a complete network, e.g., when we set $Q_{12} = Q_{21} = 0.01$, we can get a very similar grayscale plot as in Figure 2.6.

The above example shows that the complete network assumption can also be relaxed for the

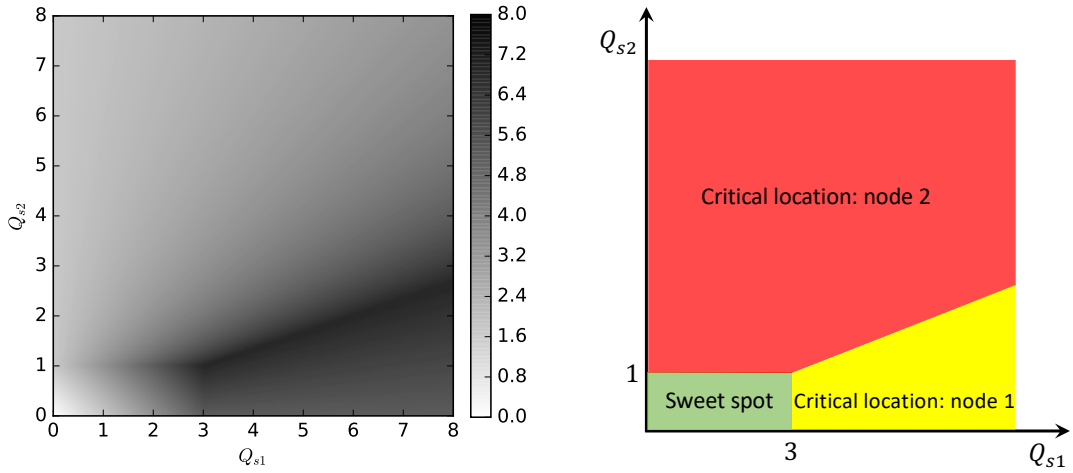


Figure 2.6: (Left) The grayscale plot of the network revenue with respect to Q_{s1} and Q_{s2}. (Right) Three regimes of parameters Q_{s1} and Q_{s2}.

paradox to occur. We can also see that the paradox is a prevalent phenomenon rather than a rare occurrence of the special case because the paradox regime is almost all the entire space (excluding the rather small "sweet spot"). Indeed, the two conditions for A_1 and A_2 in Proposition 5 are rather mild and not restrictive at all.

2.4.3 Allowable Increase

As we have seen, the critical location plays an important role in the network. We are interested in the conditions under which the critical location does not shift when the demands on arcs in the network change.

We first study the allowable increase associated with the outbound demand of the critical location j, which is defined as the amount by which we can increase the outbound demand of the critical location j while keeping the location j to be the only critical location. Suppose the outbound demand Q_{jk} is perturbed to be $Q_{jk} + \tau \cdot \varepsilon_{jk}$, for all $k \neq j$, where ϵ_j satisfies $\varepsilon_{jj} = 0$, $\sum_{l=1}^{n} \varepsilon_{jl} = 1$ and $\varepsilon_{jk} \geq 0, \forall k \neq j$. If τ is less than the critical location j's allowable increase, then the location

j remains to be the only critical location. The following result presents the lower bound for the allowable increase.

Proposition 6 (Allowable Increase)**.** *Suppose the location j is the only critical location in the network, and we increase its outbound demand Q_{jk} to be $Q_{jk} + \tau \cdot \varepsilon_{jk}$, for all $k \neq j$, where $\boldsymbol{\epsilon}_j$ satisfies $\varepsilon_{jj} = 0$, $\sum_{l=1}^{n} \varepsilon_{jl} = 1$ and $\varepsilon_{jk} \geq 0, \forall k \neq j$.*

1. *If there are only two locations in the network and the other location is location i, then the location j remains to be the only critical location as long as $0 < \tau < Q_{ij} - x_i^*$ holds.*

2. *If there are at least three locations in the network, then the location j remains to be the only critical location, as long as $0 < \tau \leq \min_{v:v\neq j} \left(\sum_{k=1}^{n} Q_{vk} - x_v^* \right) \cdot \alpha_{vj}$ holds.*

If the outbound demand of the critical location j increases by less than $\min_{v:\ v\neq j} \alpha_{vj} \cdot \left(\sum_{k=1}^{n} Q_{vk} - x_v^* \right)$, then the critical location j remains the only critical location in the network and the lower and upper bounds in Proposition 3 remain hold. Next, we study the allowable increase associated with the demand between non-critical locations.

Proposition 7 (Allowable Increase)**.** *For any complete network with at least 3 locations, suppose location j is the only critical location and the location i and k are two arbitrary non-critical locations. If the demand from i and k increases no more than the maximum $\widehat{B}_{ik}$ value that satisfies the following inequalities, then the location j remains to be the only critical location.*

$$\widehat{B}_{ik} \leq \frac{\sum_{l=1}^{n} Q_{il}}{x_i^*} \cdot \min \left\{ \sum_{l=1}^{n} Q_{il} - x_i^*,\ \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\} + \frac{\sum_{l=1}^{n} Q_{il}}{x_i^*} \cdot \frac{\alpha_{kv}}{\alpha_{iv}} \cdot \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right], \forall v \neq i,k$$

$$\widehat{B}_{ik} \leq \frac{\sum_{l=1}^{n} Q_{il}}{x_i^*} \cdot \min \left\{ \sum_{m} Q_{im} - x_i^*,\ \Phi_i \right\} + \frac{\sum_{l=1}^{n} Q_{il}}{x_i^*} \cdot \frac{\alpha_{kv}}{\alpha_{iv}} \cdot \min \left\{ \sum_{m} Q_{km} - x_k^*,\ \Phi_k \right\}$$

$$+\frac{\sum_{l=1}^{n} Q_{il}}{x_i^*}\cdot\frac{\alpha_{qv}}{\alpha_{iv}}\cdot\left(\sum_{l=1}^{n} Q_{ql}-x_q^*\right),\ \forall q\neq i,k; v\neq i,k; q\neq v$$

where

$$\Phi_i=\left(\sum_m Q_{qm}-x_q^*\right)\cdot\alpha_{qi}+\min\left\{\left(\sum_m Q_{qm}-x_q^*\right)\cdot\alpha_{qk},\sum_m Q_{km}-x_k^*\right\}\cdot\alpha_{ki}$$

$$\Phi_k=\left(\sum_m Q_{qm}-x_q^*\right)\cdot\alpha_{qk}+\min\left\{\left(\sum_m Q_{qm}-x_q^*\right)\cdot\alpha_{qi},\sum_m Q_{im}-x_i^*\right\}\cdot\alpha_{ik}$$

2.5 Intervention Strategies for the Network Operator

The paradox in Section 2.4 clearly shows the pitfall of the decentralized shared mobility system, and it highlights the necessity of the network operator's strategic intervention. Guided by above equilibrium analysis, we study two intervention strategies for the network operator to improve the network.

2.5.1 Service Region Selection

Consider the network in Figure 2.5 and assume $Q_{s1}=Q_{1s}=Q_{s2}=3$ and $Q_{2s}=1$. In the equilibrium flow, the flow on every arc is equal to 1 and the amount of the total network flow is 4. Now consider removing the location 2 from the network, and the remaining network only has location s and location 1. In the new (maximum possible) equilibrium flow, the flow from location s to location 1 and the flow from location 1 to location s are both 3; so the amount of the total network flow becomes 6. We can see that after removing the location 2, the total flow in the (maximum possible) equilibrium flow increases from 4 to 6. Even though there is no fixed cost to serve the locations, eliminating the location 2 and associated demands increases the total network flow! In the following, we present a similar result for a general hub-and-spoke network.

Consider a hub-and-spoke network with one hub at the center as location s, and n ($n > 1$) spokes around the hub as locations $1, \cdots, n$. We denote the demand from the hub to the spoke j as Q_{sj}, and from the spoke j to the hub as Q_{js}; there is no demand between any two spokes.

Because there is no setup cost with each location, it would seem that we should serve the demands of all the locations. However, we show that in the optimal policy some locations should be excluded from the service region. The hub should always be included in the service region because otherwise there would be no flow in the network. We use the binary variable d_j to denote whether the spoke j should be included in the service region: $d_j = 1$ if the spoke j is included in the service region, and $d_j = 0$ if not. The equilibrium flow can be obtained by solving the following problem.

$$\max_{\boldsymbol{x},\boldsymbol{d}} \; x_s + \sum_{j=1}^{n} d_j \cdot x_j \tag{2.5a}$$

$$s.t. \; x_j = x_s \cdot \frac{d_j \cdot Q_{sj}}{\sum_{k=1}^{n} d_k \cdot Q_{sk}}, \forall j = 1, \cdots, n \tag{2.5b}$$

$$x_s = \sum_{k=1}^{n} d_k \cdot x_k, \tag{2.5c}$$

$$x_j \leq d_j \cdot Q_{js}, \forall j = 1, \cdots, n, \tag{2.5d}$$

$$x_s \leq \sum_{j=1}^{n} d_j \cdot Q_{sj}, \tag{2.5e}$$

$$d_j \in \{0,1\}, x_s, x_j \geq 0, \forall j = 1, \cdots, n \tag{2.5f}$$

Define $r_j = \frac{Q_{js}}{Q_{sj}}$, $\forall j = 1, \cdots, n$; each ratio measures the corresponding spoke' ability to send back vehicles to the hub. Let $r_{min} = \min\{r_j : d_j = 1, \; j = 1, \cdots, n\}$. Denote the optimal solution to the problem (2.5) as $\boldsymbol{x}^*, \boldsymbol{d}^*$.

Proposition 8. *1. The optimal service region selection policy for the Model* (2.5) *has a thresh-*

old structure with respect to the ratio r_j; i.e., there exists a threshold r_0 such that $d_j = 0$ if $r_j \leq r_0$, and $d_j = 1$ if $r_j \geq r_0$.

2. *The amount of the equilibrium flow is equal to* $2 \cdot \sum_{k=1}^{n} d_k \cdot Q_{sk} \cdot \min\{r_{min}, 1\}$.

Proposition 8 clearly shows a paradoxical situation that adding a location can decrease the total network flow even though there is no fixed cost associated with each location.

If $r_{min} < 1$, then there is a spoke who has limited ability to send back vehicles to the hub, and that spoke is the critical location; if $r_{min} > 1$, then the hub is the critical location. In both cases, the critical location constrains the total network flow.

Note that the service region selection in our problem is fundamentally different from the traditional facility location problem, where closing a site is usually due to the associated fixed setup cost. In the shared mobility system, a location may still be excluded even if there is no fixed cost associated with it. The trade-off here is as follows: on one hand, adding an additional location can be beneficial because the network's total potential demand increases; on the other hand, the added location could be a critical location and the entire network's total flow could decrease.

A North American Region Director of Car2Go once told us that in their car sharing network, there are indeed areas where the cars are more likely to be stranded, which had affected the overall availability of vehicles in the entire city. These areas correspond to the critical locations in our model, and the company had to cut out such areas to ensure that the rest of the network has enough vehicles. Note that reducing the coverage area in this case was indeed strategically selecting service areas rather than scale back their business because the company Car2Go still "maintain[ed] the same number of vehicles" after excluding those areas.

2.5.2 Reposition Strategy

Because the critical location's outbound demands constrain the total flow of the entire network, the network operator may consider enhancing the outbound flows from the critical location by repositioning vehicles from it. For example, the bike sharing company BCycle deploys trucks to reposition bikes to increase the total network flow. Recall that in Proposition 4, we already proved that the network operator should prioritize repositioning vehicles from the critical location. In this section, we use a hub-and-spoke network to further study insights for the reposition strategy.

We consider a simplified hub-and-spoke network that captures key geographic features of a typical shared mobility network. Assume a location s is the hub, and there are $n > 1$ spokes denoted as locations $1, \cdots, n$. We let all the demands from the hub to spokes Q_{sj}s to be equal to 1, i.e., $Q_{sj} = 1, \forall j = 1, \cdots, n$. Let $\beta_s = \sum_{j=1}^{n} Q_{sj} = n$. We then let the demand from spokes to the hub Q_{js}s vary with each other so that different spokes have different degrees of imbalance between the inbound and outbound demand. Denote $Q_{js} = \beta_j$, and without loss of generality, assume $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$ holds. For simplicity, we assume the demands between the spokes are the same: $Q_{jk} = \theta, \forall j,k = 1, \cdots, n$, where $\theta \geq 0$; the parameter θ depicts the demand intensity between spokes.

It can be shown that only the hub or spoke 1 can be the critical location. These two cases are not fundamentally different; the major distinction between the two cases is just the shift of the critical location. Therefore, we only consider the case where the spoke 1 is the critical location. It can be shown that the condition $\sum_{j=1}^{n} \beta_j = \beta_s$ guarantees that spoke 1 is the critical location, and this condition will be assumed in this section. Note that we have assumed this condition just to simplify our analysis, but our analytical insights can be generalized to other cases because the equilibrium flows and the revenue in other cases are just scaled by a factor.

We assume the reposition takes place at the end of each period and the repositioned vehicles

will be available to use at the beginning of the next period. The cost of moving a vehicle between any location pairs is given as $c_T > 0$. Assume the price for the trip between any two locations is p_T, and the percentage of customers that are willing to pay is $\big(1 - F(p_T)\big)$. Denote the amount of vehicles repositioned from location i to location j in each period as z_{ij}. Assume the reposition capacity is K. The revenue maximizing problem with vehicle reposition is as follows:

$$\max_{p_T, \boldsymbol{x}, \boldsymbol{z}} \; p_T \cdot x_s + p_T \cdot \sum_{j=1}^{n} x_j - c_T \cdot \sum_{i,j} z_{ij} \tag{2.6a}$$

$$s.t.\; x_j = \frac{x_s}{n} + \sum_{k=1, k\neq j}^{n} x_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} + z_{sj} + \sum_{k=1, k\neq j}^{n} z_{kj} - z_{js} - \sum_{k=1, k\neq j}^{n} z_{jk}, \forall j = 1, \cdots, n \tag{2.6b}$$

$$x_s = \sum_{k=1}^{n} x_k \cdot \frac{\beta_k}{\beta_k + (n-1)\cdot\theta} + \sum_{k=1}^{n} z_{ks} - \sum_{k=1}^{n} z_{sk}, \tag{2.6c}$$

$$x_j \leq \big(\beta_j + (n-1)\cdot\theta\big) \cdot (1 - F(p_T)), \forall j = 1, \cdots, n, \tag{2.6d}$$

$$x_s \leq \beta_s \cdot (1 - F(p_T)), \tag{2.6e}$$

$$\sum_{k=1}^{n} \left(z_{js} + \sum_{k=1, k\neq j}^{n} z_{jk} \right) \leq K, \tag{2.6f}$$

$$x_s, x_j, z_{sj}, z_{js}, z_{jk} \geq 0, \forall j, k = 1, \cdots, n \tag{2.6g}$$

Let $\Delta_s = \sum_{k=1}^{n} z_{sk} - \sum_{k=1}^{n} z_{ks}$, and $\Delta_j = z_{js} + \sum_{k=1, k\neq j}^{n} z_{jk} - z_{sj} - \sum_{k=1, k\neq j}^{n} z_{kj}, \forall j = 1, \cdots, n$. Positive (negative) Δ value means there are vehicles repositioned out from (into, respectively) the corresponding location.

Proposition 9. *There exist* j_1, j_2 $(0 \leq j_1 < j_2 \leq J+1)$ *such that the optimal reposition policy*

satisfies the following threshold structure:

$$\Delta_j \begin{cases} > 0, & j = 1, \cdots, j_1, \\ = 0, & j = j_1 + 1, \cdots, j_2 - 1, \\ < 0, & j = j_2, \cdots, n, \end{cases} \tag{2.7}$$

Under the condition $\sum_{j=1}^{n} \beta_j = \beta_s$, the hub would never be a critical location because its outbound and inbound demands are balanced. Because we have assumed that $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$, the spoke 1 is the critical location if the network operator does not reposition vehicles. When the reposition capacity K is very small, the network operator should prioritize repositioning vehicles from the critical location because it has the least ability to send out vehicles. Therefore, the reposition should start with the critical location – spoke 1. As K increases, more vehicles can be repositioned from spoke 1. At some point, spoke 2 would become a new critical location. Then the network operator should reposition vehicles from both spoke 1 and spoke 2, i.e., j_1 becomes 2. As K continues to increase, j_1 would increase, and more spokes would become critical locations. Consequently, the network operator should reposition their stranded vehicles as well.

Spoke n has more outbound demands than inbound demands, so if vehicles from other spokes are repositioned to spoke n, more outbound demands from spoke n can be satisfied. As the reposition capacity K grows large, more and more vehicles are repositioned to spoke n. After enough vehicles are repositioned to spoke n, the network operator should reposition vehicles to spoke $n-1$, then to $n-2$, $n-3$ and so on.

2.6 Conclusion

We use the proportional vehicle movement to model the decentralization of customer flows in the shared mobility network. We show that no matter how the system is initialized, the network flows can reach the globally stable equilibrium without the network operator's intervention. We develop a linear model to characterize the equilibrium flow. Based on the optimal solution of the linear model, we identify a critical location whose total outbound demands constrain the total network flows. We show that in the equilibrium flow, there can be idle vehicles at the critical location while all other locations have unsatisfied demands. We then demonstrate a paradox regarding the critical location and we also show that increasing the critical location's outbound demand has a multiplier effect on the total network flow. These analytical results lead to two intervention strategies for the network operator to improve the network. The network operator may need to exclude certain locations from the service region even if there is no fixed cost to serve the locations. The network operator can also reposition vehicles to increase the network flows, and we show that the vehicles at the critical location should be prioritized for reposition.

In this study, we have assumed that the demands are time-invariant. However, this may not be realistic for some real-world applications. For example, on workdays, bikes in the bike-sharing network come into the downtown area in the morning and disperse into surrounding residential neighborhoods in the afternoon. We can extend our model to incorporate such cyclical seasonal demand pattern, and analyze the impact of the temporal demand variation on the network operations.

Chapter 3

Extended Open Shop Scheduling with Resource Constraints: Appointment Scheduling for Integrated Practice Units [1]

3.1 Introduction

The United States spent nearly 18% of its gross domestic product on healthcare in 2015 according to the Centers for Medicare & Medicaid Services (CMS 2018). In 2016, U.S. healthcare spending reached a new peak at $10,348 per person, more than twice the average of other developed countries. Today, it is most common for patients who need multiple consults to travel from one clinic to another to see different providers. Such a provider-centered approach inevitably burdens the patient in the following ways: (1) travel between facilities introduces inconvenience, additional logistics costs and unnecessary administrative costs; (2) repeated requests for the same information can increase stress and anxiety; (3) information transfer across clinics often results in inaccurate or incomplete health records downstream; (4) lack of communication among providers may occasion unnecessary or duplicate tests, and undermine long-term care planning; and (5) the separation of providers reinforces a piecemeal approach that rarely addresses the patient's condition as a whole. To better deliver healthcare services, current healthcare reform is moving towards value-based patient-centered care, seeking better coordination among providers. Many researchers have shown that this approach can improve clinical outcomes while decreasing diagnostic tests and the need for referrals (Hanna 2010; Stewart et al. 2000).

[1] This chapter is based on Zhang et al. (2019). I appreciate Prof. Douglas Morrice and Prof. Jonathan Bard for their support and guidance for this paper.

3.1.1 Integrated Practice Units

To put the focus on the needs of the patient, several clinicians and policy analysts have suggested the use of integrated practice units (IPUs) to treat chronic medical conditions such as diabetes, pain, multiple sclerosis, and cardiomyopathy, to name a few (Porter 2010; Keswani et al. 2016). This approach fosters realtime communication among specialists and provides treatment options for the patient across the entire continuum of care for a chronic condition. An added benefit inherent to this model is the continuous learning and improvement of multiple disciplines working together and communicating about each patient. The team learns from every patient so their expertise improves over time. After a patient enters the IPU and is roomed, the appropriate providers sequentially address the patient's conditions. In some cases, it is appropriate for the patient to see different providers in a specific order. For example, in the case of a lower extremity joint pain IPU, the motivating clinic for this paper, the patient is first seen by a nurse practitioner who determines whether additional treatment is required. If it is decided that the patient needs to see both a surgeon and a physical therapist, the surgeon comes first. If it is determined that the patient must see a physical therapist and a nutritionist, the order is immaterial.

What is relatively unique about an IPU is that the patient remains in the same room for the duration of his visit, and hence is the center of pathways traversed by a variety of providers. This model of care delivery enables the providers to work more closely together in treating their patients, and to focus on using the skills for which they have been trained. The expectation is better outcomes, higher levels of patient satisfaction, and lower patient costs in the long run. What has yet to be determined, though, is whether the efficiency of an IPU will outweigh the higher provider costs that are likely to result from lower provider utilization. To be effective, all providers must be available in the IPU but not all patients need to see all providers.

While IPUs bring continuity of care and integrated treatment to patients – important factors

in patient satisfaction – they also present an operational challenge to the schedulers who must coordinate activities among all providers. In the current system, healthcare delivery is fragmented; patients see their providers at different times and often at different locations. In an IPU, the patients have seamless access to service from different providers in a timely manner; however, this requires better coordination among providers to prevent delays and congestion. Moreover, in the fragmented delivery system, different clinics operate independently and each has its own scheduling system. In an IPU, the schedules of the different providers interact with each other because the patient needs to see one provider followed by another. As a consequence, clinic scheduling becomes central to the efficiency of the multidisciplinary team because patient demand and provider capacity have to be strategically matched to ensure timely operations.

One of the most critical issues in managing an IPU is deriving the appointment schedule or template. Ill-conceived templates result in excessive patient waiting time, unacceptably low provider utilization, and costly overtime. The challenge then is to design schedules that jointly balance clinic closing time and total patient waiting time while also taking into account system capacity and system randomness. For a given number of patients, if these two metrics are minimized, then provider and staff idle time should also be minimized. The system randomness derives from two sources. The first is each patient's provider set. These sets are unknown at the time when the appointment is made and are only determined after the patient is seen by the nurse practitioner who conducts an initial examination. The second is the amount of time each patient spends with each provider.

In practice, the coordination of providers in IPUs has many elements of an extended open shop scheduling problem in which each workstation may consist of multiple (identical) machines and some jobs have partially fixed routes. The pure open shop problem has been studied extensively in the combinatorial optimization literature, and is known to be strongly NP-hard (see Pinedo 2016).

Its aim is to assign a set of jobs to available machines to minimize one of several objective functions such as makespan, total processing time, or number of late jobs. Because the IPU appointment scheduling problem has many similar characteristics as the open shop problem, the models for either are quite similar. In the case of an IPU, the patients can be viewed as jobs and the providers as machines. Clinic performance is measured by patient delay (total processing time) and closing time (makespan). These measures are in conflict so a strategic balance must be struck.

3.1.2 Research Contributions

The purpose of this paper is to first present a generic model of the IPU scheduling problem and then to develop a solution methodology for realistic size instances. We focus on two decisions: the number of patients to schedule in each time period and appointment rules. The first decision is intended to fix the appointment template, which specifies how many patients should be scheduled to arrive at the beginning of each time slot. Appointment rules determine which types of patients (new or follow-up) to assign to each time slot. We begin with a deterministic model based on an open shop that takes into account the unique characteristics of an IPU including different types of providers, multiple providers of the same type, fixed and variable patient paths, and patient waiting time limits. An additional consideration is the number of available rooms. Once a patient checks into the clinic, he is assigned to a room and remains there until the visit is concluded. This type of resource constraint is not often modeled in open shop scheduling problems where the common restrictions center on labor and machines. It is rare for auxiliary resources, such as rooms, transportation vehicles, and other tooling and equipment to be taken into account. We propose three approaches to modeling such resource constraints. To capture the randomness of provider service times and patient-specific treatments, we show how our generic model can be extended to include these stochastic elements. The approach is demonstrated using data provided

by The University of Texas Dell Medical School in Austin.

This research differs from earlier studies in the following ways. To the best of our knowledge, this is the first paper to present a generic (stochastic) model for determining appointment templates for a multi-stage, multi-server, resource constrained clinic where patients remain stationary throughout their visit. Another unique feature of our problem is the order of provider-patient engagement. Existing studies usually assume that the patient sees the providers in a fixed sequence if there is more than one, while in our case, the order is only fixed for some providers while remaining flexible for the others. Thus, the IPU scheduling problem is really a combination of a flexible flow shop and an open shop with auxiliary resource constraints (see Pinedo 2016). The two-step method proposed to find solutions is sufficiently general to be used to help solve similar coordinated appointment scheduling problems arising from other applications.

The remainder of this paper is structured as follows. In Section 3.2, we provide a literature review of the most relevant work on open shop scheduling and healthcare appointment scheduling. In Section 3.3, we present our new model for the extended open shop scheduling problem, analyze its features, and introduce several valid inequalities that were seen to speed convergence. We also describe our two-step solution method. The random components of the problem are introduced in Section 3.4 where we present a two-stage stochastic optimization model and define what we mean by the expected value solution and the wait-and-see solution. In Section 3.5, we examine the relative performance of two templates derived from the expected value solution and two found in the literature. Extensive testing is done to compare IPU metrics across all templates and to evaluate the quality of the lower bound obtained from the two-step method. The results indicate that the average gap for the two-step method is always less than 5% for the wait-and-see problem, and less than 2% for the four appointment templates that we investigated.

We also observed that the two templates derived from the expected value solution are good

candidates for setting appointments. One template emphasizes the clinic closing time objective by scheduling patients to arrive relatively earlier in the day. The second template emphasizes the patient waiting time objective by scheduling patients to arrive later in the day. Lastly, the results show that our appointment rules are helpful when scheduling the different types of patients. For example, we found that it is best to schedule follow-up patients, who generally have shorter service times, to arrive when there is high patient flow. This helps to relieve or avoid congestion when the number of patients is fixed over the day. We conclude with some managerial insights and some suggestions for future research in Section 3.6.

3.2 Literature Review

Variations of job shop problems have been studied extensively and have a wide range of applications. One common example is the open shop problem in which a set of jobs is to be processed through multiple stations in an arbitrary order, as is partially the case in an IPU. Bhat et al. (2000) modeled the communication scheduling problem as an open job shop while Liaw (2000) proposed a hybrid genetic algorithm that incorporated tabu search as part of the solution methodology. Noori-Darvish et al. (2012) developed a bi-objective mixed-integer linear programming (MILP) model for an open shop scheduling problem with sequence-dependent setup times, and applied an interactive fuzzy programming approach to find solutions. Our clinic scheduling problem can be modeled as an extended open shop, where "extended" means multiple, parallel machines, fixed and arbitrary job processing paths, and auxiliary resource constraints.

Scheduling problems in healthcare often have special features that distinguish them from problems arising in other industries. Their unique nature brings additional challenges. For example, Gupta and Denton (2008) note that in healthcare applications there exists less flexibility because patients may have a preference for a specific provider or appointment time. Moreover, urgent pa-

tient needs must be accommodated immediately, and in some cases, price cannot be used to modulate patient demand. With respect to outpatient scheduling, a wide variety of approaches have been investigated but few have been implemented in practice. In the remainder of this section, we provide a literature review of healthcare scheduling problems with different system structures. We also provide a review of the different solution methods with an emphasis on stochastic programming approaches.

3.2.1 Healthcare Systems with Different Pathway Structures

There have been numerous studies on scheduling in healthcare over last several decades, as highlighted by Cayirli and Veral (2003) and Gupta and Denton (2008). Most of the early work focused on single-station appointment scheduling. More recently, the scope has expanded to include multi-stage, multi-server applications as discussed by Ahmadi-Javid et al. (2017) and Leeftink et al. (2018). Based on the features of our problem, the most relevant studies can be grouped into two categories: multi-stage models and multi-server models. Each is reviewed below.

In multi-stage clinic scheduling, different provider types are involved. This makes the problem complicated because a patient can be referred from one provider to another for different treatment, which leads to uncertainty in the patient flow. Azadeh et al. (2015) formulated a semi-online patient scheduling problem as a MILP, and developed a genetic algorithm to find solutions. In their problem, the patients require different types of tests and the use of a variety of laboratory equipment. Castro and Petrovic (2012) studied a scheduling problem in which patients need to go through an ordered sequence of examinations. They formulated the problem as a three-objective mathematical program, and solved it with a dispatching rule. Pérez et al. (2013) investigated a stochastic online scheduling problem for nuclear medicine clinics where the patients need to go through multiple steps. In the study, the sequence of the steps is fixed, and multiple resources are

required at each step. Kazemian et al. (2017) developed a simulation model to coordinate clinic and surgery appointments with the objective of reducing the indirect waiting time of patients and limiting operating room overtime. Their strategy was to choose appointment days for patients rather than setting daily arrival times. Different from our work, these studies are either limited to a single server at each stage or they do not include room constraints.

Problems get more challenging when there is more than one provider of each type, giving rise to the multi-server clinic scheduling problem. Gupta and Wang (2008) modeled an appointment booking problem as a Markov decision process and proposed heuristics to find solutions. They also developed lower and upper bounds on the optimal solution, which were shown to speed convergence. Both single- and multiple-physician clinics were analyzed, but in either case, only single-stage scheduling was applicable. Parizi and Ghate (2016) went a step further and purposed a Markov decision process for a multi-class, multi-resource clinic scheduling problem, while Qu et al. (2013) developed a weekly scheduling template for a multiple-provider outpatient clinic. In their problem, providers in separate sessions have separate appointment schedules, while in our study, all providers are in the same clinic working with a single appointment schedule.

3.2.2 Solution Methods for Healthcare Scheduling Problems

Dynamic programming has been a popular tool for modeling the clinic scheduling problem. For example, Truong (2015) considered the problem in which two types of patients are adaptively given appointments over several days. Chakraborty et al. (2010) used a dynamic programming tree to investigate clinic scheduling with general service time distributions where the patients sequentially request appointments. Simulation is perhaps the most versatile tool since it is able to handle most complexities surrounding patient flow and uncertainty. Wang et al. (2018) solved a two-server scheduling problem using simulation-based optimization. Cayirli et al. (2006) devel-

oped a simulation model to analyze appointment scheduling for ambulatory care and investigated patient sequence rules based on patient class. Similarly, Bard et al. (2016) used discrete event simulation to investigate the performance of the family health center associated with the University of Texas Medical School in San Antonio. Their objective was to obtain a better understanding of patient flow and to evaluate changes to current scheduling rules and operating procedures. As part of the study, they examined a variety of scenarios related to appointment scheduling and managing early and late arrivals.

Robust optimization is a relatively new approach to scheduling patients and resources in healthcare facilities. Denton et al. (2010) built a robust optimization model to study the allocation of operating rooms to surgical specialties in the face of insufficient data. Rachuba and Werners (2014) applied the robust approach to a hospital surgery scheduling problem in an effort to avoid frequent rescheduling due to random requests and cancellations. Similarly, Mannino et al. (2012) presented a light robustness procedure to handle random fluctuations in demand when constructing cyclic master surgery schedules. In their procedure, parameter values lie in an uncertainty set but solutions are not required to satisfy all possible realizations. Instead, soft constraints are introduced for each parameter and violations are penalized in the model's objective function.

Another common approach to modeling uncertainty is stochastic programming. Mancilla and Storer (2012) considered a stochastic appointment scheduling problem and proposed a new sequencing algorithm based on Benders decomposition to find solutions. Oh et al. (2013) used a stochastic integer programming model to schedule patient appointments in primary care facilities and developed scheduling guidelines. Integral to their work is (i) an empirically based classification scheme to distinguish chronic and acute conditions, (ii) the ability to coordinate patient and provider interactions, and (iii) the introduction of slack in the schedule to accommodate the effects of service time variability. Kong et al. (2013) investigated an outpatient clinic appointment

scheduling problem with a single physician and proposed a convex conic programming approach to find solutions. Berg et al. (2014) considered a profit-maximization scheduling problem in the presence of patient no-shows and random procedure times. They modeled the problem as a two-stage stochastic mixed-integer program and proposed several methods to find solutions including two decomposition approaches and a heuristic.

Chen and Robinson (2014) formulated a clinic scheduling problem with both routine patients and last-minute patients as a stochastic linear program. They derived optimal sequencing rules while accounting for random no-shows and call-ins. Erdogan and Denton (2013) proposed a multi-stage stochastic linear program in which each stage is defined to coincide with the time a patient calls to request an appointment. Different from the formulations in these studies, our two-stage optimization model accounts for resources shared among patients and co-located providers who see patients in a partially fixed and partially random order.

3.3 Deterministic Model

As noted in Section 3.1.1, it is critical to consider uncertainty when designing appointment templates for IPUs. The foundation of our approach is a stochastic optimization model whose solution relies heavily on efficiently solving a deterministic version of an extended open shop scheduling (EOSS) problem. In Section 3.3.1, we present our EOSS model that includes parallel machines at each station. After describing the formulation, we highlight its unique features in Section 3.3.2 and offer some tightening constraints designed to reduce the computational burden. To make the discussion concrete, the focus is on clinic scheduling, but with the understanding that the model is generally applicable to most open shop problems. Next, in Section 3.3.3 the formulation for the room constraints is presented. These constraints can readily handle similar resources such as vehicles, jigs, tooling, and auxiliary personnel. In Section 3.3.4 we specialize the open shop model

to an IPU and impose additional restrictions that better reflect operational considerations. Finally, in Section 3.3.5 we propose a two-step heuristic to obtain upper and lower bounds on the optimal schedule.

3.3.1 Extended Model for Open Shop Scheduling

We first study the general minimum makespan extended open shop scheduling problem with a secondary objective of minimizing the total time that jobs spend in the system. The presentation reflects clinic appointment scheduling rather than job shop scheduling. In the developments, we make use of the following notation.

Indices and sets

i, j	index for patients
k, l	index for providers or provider types
m	index for position in the sequence of patients who see a particular provide type
o	origin (and destination) index for all patients and all providers
J	set of patients
K	set of provider types
$J(k)$	set of patients who see type k provider
$K(j)$	set of provider types that patient j needs to see

Data and parameters

adm^k	time (hours) required for a type k provider to perform administrative functions such as entering data into the electronic medical records system after seeing each patient

LT_m^k lower bound on $m+1^{\text{st}}$ patient's starting time with a type k provider. (When there is only one provider of type k, LT_m^k equals the sum of the m smallest service times of provider k's patients. It is also the lower bound on the time interval between any two patients who are separated by $m-1$ other patients.)

m^k total number of patients that type k providers are to see

n^k number of type k providers

s_j^k service time required for a type k provider to treat patient j (hours)

ε ratio of the predetermined waiting time that a patient can spend in clinic to the patient's total service time

$S_j(\varepsilon)$ upper limit on the amount of time that patient j is allowed to spend in the clinic, or equivalently, the total service time plus upper limit on waiting time of patient j; that is $(1+\varepsilon)\cdot\sum_{k\in K(j)} s_j^k$

T_{max} upper bound on clinic closing time

Decision variables

t_m^k start time of the patient in the m^{th} position in the schedule of type k providers

x_{jm}^k 1 if patient j is in the m^{th} position in the sequence of patients who see a type k provider, 0 otherwise

ST_j^k time when a type k provider starts seeing patient j

y_j^{kl} 1 if $ST_j^k + s_j^k \le ST_j^l$, which means that a type k provider must finish his visit with patient j before a type l provider can start seeing patient j;

0 if $ST_j^k \ge ST_j^l + s_j^l$, which means that a type k provider can start seeing patient j no earlier than a type l provider finishes his visit with patient j

Accounting variables

T	clinic closing time
T_j^1	time when patient j is seen by his first provider
T_j^2	time when patient j finishes being seen by his last provider

For the clinic scheduling problem, we are given a set J of $|J|$ patients and a set K of $|K|$ provider types. For each $k \in K$ there are n^k providers. Different providers of the same type can perform the same tasks. Each patient $j \in J$ needs to be seen by a subset of providers, denoted by $K(j)$. The service time for patient j when treated by a type k provider is s_j^k. As in the general open shop scheduling model, there is no restriction on the order in which providers can see patients.

Each patient is visited by one provider at a time and cannot be preempted once service begins. When the provider finishes treating a patient, she documents the episode. This requires a moderate amount of administrative time but does not affect the patient who can be seen immediately by another provider. The objective is to minimize a weighted combination of the makespan (clinic closing time) and the patients' total time in clinic. The makespan is our primary concern, and in the implementation, is assigned a much larger weight than the total time patients spend in the facility.

To simplify the presentation, first consider the case where $n^k = 1$ for all $k \in K$, where the m^k patients to be seen by the type k provider are indexed by m (i.e., $m = 1, 2, \ldots, m^k$). The decision variable x_{jm}^k is associated with patient $j \in J(k)$ and takes the value of 1 if patient j is in position m in provider k's schedule, and 0 otherwise. The benefit of this indexing scheme is that if a position has a lower/higher index, then the starting time associated with this position should also be lower/higher. Accordingly, the position index can be used to calculate lower and upper bounds on the starting time of the corresponding patient.

Now consider the case where $n^k > 1$. For any k, a corresponding provider can see at most m^k patients. Therefore, we need at most $n^k \cdot m^k$ binary x-variables for each $j \in J(k)$ to determine which of the m^k providers treats patient j, as well as the order in which patients are seen. To help formulate the constraints, we put each provider's patient positions into different sets. Figure 3.1 depicts an example with three providers A, B and C of the same type. In the model, there are $3 \cdot m^k$ positions indexed as $1, 2, \ldots, 3 \cdot m^k$, where each position is marked as A, B or C. The patients who are assigned the positions marked with an A (B or C), will be seen by provider A (B or C, respectively). In the example, provider A's patients will be in positions 1, 4, 7, ... Since we have m^k patients and $3 \cdot m^k$ positions, only m^k positions will be filled by the patients in a solution; the remaining $2 \cdot m^k$ positions will be empty.

$$\underbrace{\text{A, B, C}}_{\text{first set}}, \underbrace{\text{A, B, C}}_{\text{second set}}, \text{A, } \ldots \text{ , } \underbrace{\text{A, B, C}}_{m^k th \text{ set}}$$

Figure 3.1: Patient positions for provider type k with 3 providers

For the general case with n^k type k providers and m^k patients, we have m^k sets, with each set containing n^k positions. The first patient in each set is seen by the first type k provider, the second patient is seen by the second type k provider, and so on. The n^{th} patient in the m^{th} set is the m^{th} patient seen by the n^{th} type k provider. In a solution, only m^k out of the $n^k \cdot m^k$ positions will be occupied. For provider type k, the binary variable x^k_{jm} specifies which position patient j takes, and according to the indexing scheme, the value of m determines which provider the patient sees. In a preprocessing step it is possible to eliminate a large number of the m^k variables associated with type k providers when $n^k > 1$. This is a direct consequence of the following assumption concerning provider-patient assignments.

In the model, we assume without loss of generality that the number of patients assigned to

providers of the same type is non-increasing. If there are three type k providers, for example, and 21 ($= n^k$) patients, then the first provider can see up to 21 patients, the second provider can see a maximum of 10 patients, and the third provider can see a maximum of 7 patients. A second benefit of the position indexing scheme is that it allows for the implementation of this ordering rule in a straightforward manner.

The model for the EOSS problem is as follows.

$$\min \alpha_1 \cdot T + \alpha_2 \cdot \sum_{j \in J} (T_j^2 - T_j^1) \tag{3.1a}$$

$$s.t. \sum_{1 \le m \le n^k \cdot m^k} x_{jm}^k = 1, \qquad \forall\ j \in J,\ k \in K(j) \tag{3.1b}$$

$$\sum_{j \in J(k)} x_{jm}^k \le 1, \qquad m = 1, \ldots, n^k \cdot m^k, \forall\ k \in K \tag{3.1c}$$

$$\sum_{j \in J(k)} x_{jm}^k \le \sum_{j \in J(k)} x_{j,m-n^k}^k, \qquad m = n^k + 1, \ldots, n^k \cdot m^k, \forall\ k \in K \tag{3.1d}$$

$$\sum_{j \in J(k)} x_{jm}^k \ge \sum_{j \in J(k)} x_{j,m+1}^k,$$

$$m \in \{1, 2, \ldots, m^k \cdot n^k\} \setminus \{n^k, 2 \cdot n^k, \ldots, m^k \cdot n^k\}, \forall\ k \in K \tag{3.1e}$$

$$t_m^k - t_{m-n^k}^k \ge \sum_{j \in J(k)} x_{j,m-n^k}^k \cdot (s_j^k + adm^k),$$

$$m = n^k + 1, \ldots, n^k \cdot m^k, \forall\ k \in K \tag{3.1f}$$

$$y_j^{kl} + y_j^{lk} = 1, \qquad \forall\ k \ne l, k, l \in K(j) \tag{3.1g}$$

$$ST_j^l \ge ST_j^k + s_j^k - (1 - y_j^{kl}) \cdot S_j(\varepsilon), \qquad \forall\ j \in J, \forall\ k \ne l, k, l \in K(j) \tag{3.1h}$$

$$ST_j^k \le t_{m \cdot n^k + n}^k + (1 - x_{j,m \cdot n^k + n}^k) \cdot T_{max},$$

$$m = 0, \ldots, m^k - 1, n = 1, \ldots, n^k, \forall j \in J, k \in K(j) \tag{3.1i}$$

$$ST_j^k \ge t_{m \cdot n^k + n}^k - (1 - x_{j,m \cdot n^k + n}^k) \cdot T_{max},$$

$$m = 0, \ldots, m^k - 1, n = 1, \ldots, n^k, \forall j \in J, k \in K(j) \tag{3.1j}$$

$$T_j^1 \leq ST_j^k, \qquad \forall\ j \in J, k \in K(j) \tag{3.1k}$$

$$T_j^2 \geq ST_j^k + s_j^k, \qquad \forall\ j \in J, k \in K(j) \tag{3.1l}$$

$$T_j^2 - T_j^1 \leq S_j(\varepsilon), \qquad \forall\ j \in J \tag{3.1m}$$

$$t^k_{n^k \cdot m^k - n} + \sum_{j \in J(k)} x^k_{j, n^k \cdot m^k - n} \cdot (s_j^k + adm^k) \leq T,$$

$$n = 0, 1, \ldots, n^k - 1, \forall\ k \in K \tag{3.1n}$$

$$x_{jm}^k, y_j^{kl} \in \{0,1\}, T, t_m^k, ST_j^k, T_j^1, T_j^2 \geq 0,$$

$$\forall\ i, j \in J,\ m = 1, \ldots, n^k \cdot m^k, k \neq l, k, l \in K \tag{3.1o}$$

The objective function (3.1a) minimizes the weighted sum of the clinic closing time and the total time patients spend in treatment (check-in and rooming can be ignored because they are assumed to take a constant amount of time; they are omitted for simplicity). The weights α_1 and α_2 should be chosen to reflect the relative importance of each term. In the application, the first term dominates the second, which means that the closing time should be made as small before minimizing the total time in the system. To meet this objective, we set $\alpha_1 \gg \alpha_2$.

Constraints (3.1b) ensure that every patient j will be seen by exactly one provider of each type in his provider set $K(j)$. Note that (3.1b) is a collection of mutually disjoint special ordered set (SOS) constraints. In each constraint associated with the (j,k) pair, only one x variable will be 1 and all others 0. Exploiting this structure in the implementation greatly reduced the computational effort.

Constraints (3.1c) guarantee that every position in provider type k's schedule is assigned to at most one patient. Constraints (3.1d) ensure that for each type k provider, positions are assigned in increasing order, starting with 1 and going up to $n^k \cdot m^k$. When $n^k = 1$, all m^k positions will be

filled. When $m^k > 1$, each provider has m^k available positions but not all of them will be assigned. Although it seems that this could result in multiple optimal solutions, because the positions are assigned in numerical order this will never be the case. Constraints (3.1e) specify that if there is more than one provider of type k, then the first provider is always assigned at least as many patients as the second, the second at least as many as the third, and so on. This rule also prevents multiple optimal solutions and has the added benefit of removing symmetry among providers of the same type.

Constraints (3.1f) specify that for a provider of type k, every patient assigned to her needs to be separated in time by at least the service time of the patient in the prior position plus the administrative time (there are no constraints for the first n^k positions because they are occupied by the first patient of the n^k providers). This ensures that providers have enough time between two successive patients. Constraints (3.1g) are written only for those patients who are to be seen by providers k and l, and enforce the condition that the visits take place in sequence. Constraints (3.1h) ensure that a provider can only start a visit with a patient after the prior provider finishes with the patient.

Constraints (3.1i) and (3.1j) define patient j's starting time with each provider type while constraints (3.1k) ensure that the clinic visit for patient j begins no later than the time when he sees any of his providers. Constraints (3.1l) guarantee that the ending time of patient j's visit is no earlier than the time when he sees any of his providers plus the corresponding service time. Constraints (3.1m) limit the total time patient j spends in the clinic (total service time plus total waiting time) to be no greater than a threshold $S_j(\varepsilon)$ proportional to his total service time. Although the second term in the objective function is aimed at minimizing total clinic time, constraints (3.1m) are not redundant. Without these constraints, some patients may spend an excessive amount of time in the clinic – a result that we wish to avoid.

Constraints (3.1n) indirectly define the clinic closing time by restricting it to be no earlier than the ending times of all providers. Alternatively, we could have defined the closing time as the time when the last patient leaves, but in the linear programming (LP) relaxation, this value is much smaller than the providers' ending times due to the weakness of constraints (3.1i) and (3.1j). Using the proposed definition led to tighter LP relaxations and shorter runtimes. Finally, all variables are defined in constraints (3.1o).

3.3.2 Model Analysis and Improvement

In this section, we investigate some of the characteristics of model (3.1a) - (3.1o). First we show how to use the index information associated with each position to improve the formulation. Next, we show how the LP relaxation can be tightened.

3.3.2.1 Index Information and Valid Inequalities

The index information for two patients seen by the same provider indicates their relative order. Consider provider type k with $n^k = 1$ and m^k patients. The index of the first patient position is 1 and all other positions for that provider have a later starting time. Given that the positions are ordered, and any two successive positions are separated by the first patient's service time plus the provider's administrative time, we can derive lower and upper bounds on the starting time of each position using its index. For example, the lower bound on the starting time of the second patient is the smallest service time of all patients that are seen by provider k plus his administrative time, which is denoted by LT_1^k. The upper bound on the starting time of the last patient position is T_{max} minus the smallest service time of all patients seen by provider k plus his administrative time, denoted by $T_{max} - LT_1^k$. Generally, for provider k with $n^k = 1$, the lower bound on the starting time of provider k's patient in position m is LT_{m-1}^k and the upper bound is $T_{max} - LT_{m^k-m+1}^k$.

These bounds allow us to strengthen constraints (3.1i) and (3.1j). In the LP relaxation of model (1), (3.1i) and (3.1j) are weak constraints due to the need to make T_{max} sufficiently large to avoid cutting off any feasible solutions. As a consequence, the relaxed feasible region is too large for branch and bound to be effective for instances of realistic size. It will be seen, however, that replacing (3.1i) with (3.2a) and (3.2b), and (3.1j) with (3.2c) and (3.2d) provides a tighter LP relaxation. Note that (3.2a) and (3.2c) are for $n^k = 1$, and (3.2b) and (3.2d) are for $n^k > 1$. 3

$$ST_j^k \leq t_{m+n}^k - \sum_{m' \leq m-1} LT_{m-m'}^k \cdot x_{j,m'+n}^k + \sum_{m' \geq m+1} (T_{max} - LT_{m^k-m'+m}^k) \cdot x_{j,m'+n}^k,$$

$$m = 0, \ldots, m^k - 1,\ n = n^k = 1,\ \forall j \in J, \forall\ k \in K(j) \tag{3.2a}$$

$$ST_j^k \leq t_{m \cdot n^k+n}^k - \sum_{m' \leq m-1} LT_{m-m'}^k \cdot x_{j,m' \cdot n^k+n}^k + \left(1 - \sum_{m' \leq m} x_{j,m' \cdot n^k+n}^k\right) \cdot (T_{max} - LT_{m+1}^k),$$

$$m = 0, \ldots, m^k - 1,\ n = 1, \ldots, n^k, n^k > 1,\ \forall j \in J, \forall\ k \in K(j) \tag{3.2b}$$

$$ST_j^k \geq t_{m+n}^k - \sum_{m' \leq m-1} (T_{max} - LT_{m^k-m+m'}^k) \cdot x_{j,m'+n}^k + \sum_{m' \geq m+1} LT_{m'-m}^k \cdot x_{j,m'+n}^k,$$

$$m = 0, \ldots, m^k - 1,\ n = n^k = 1,\ \forall j \in J, \forall\ k \in K(j) \tag{3.2c}$$

$$ST_j^k \geq t_{m \cdot n^k+n}^k - \left(1 - \sum_{m' \geq m} x_{j,m' \cdot n^k+n}^k\right) \cdot (T_{max} - LT_{m^k-m-1}^k) + \sum_{m' \geq m+1} LT_{m'-m}^k \cdot x_{j,m' \cdot n^k+n}^k,$$

$$m = 0, \ldots, m^k - 1,\ n = 1, \ldots, n^k, n^k > 1,\ \forall j \in J, \forall\ k \in K(j) \tag{3.2d}$$

Proposition 10. *Collectively, constraints* (3.2a) *and* (3.2b) [*constraints* (3.2c) *and* (3.2d)] *are stronger than their counterparts constraints* (3.1i) [*constraints* (3.1j)].

The inequalities in the proof show the tightness of the improved constraints (3.2) given their

equivalence to the original two constraints (3.1i) and (3.1j). As noted, the index formulation is a unique feature of our model and is useful in tightening constraints and breaking symmetry. These advantages are not available with the more traditional routing formulation in which the subscripts on the x variables represent the immediate sequence of two entities, such as vehicles, jobs or patients. In our computational testing, we found that the tightened constraints greatly reduced runtimes.

3.3.2.2 Linear Programming Relaxation

Tight LP relaxations of MILPs are essential for computational efficiency. In model (3.1), this is partially achieved with constraints (3.1f), which enforce a minimum separation time between patients who are on the schedule of the same provider. To see this, we sum constraints (3.1f) for a single type k provider. Assume that $n^k = 1$ and denote provider k's ending time by $t^k_{m^k+1}$. This leads to

$$\begin{aligned} t^k_{m^k+1} - t^k_1 = \sum_{m=2}^{m^k+1} t^k_m - t^k_{m-1} &\geq \sum_{m=2}^{m^k+1} \sum_{j \in J(k)} x^k_{j,m-1} \cdot (s^k_j + adm^k) \\ &\geq \sum_{j \in J(k)} \left(\sum_{m=2}^{m^k+1} x^k_{j,m-1} \right) \cdot (s^k_j + adm^k) \\ &= \sum_{j \in J(k)} (s^k_j + adm^k) \end{aligned}$$

which shows that a provider's ending time and starting time are separated by at least his patient's total service time and administrative time. Considering that our primary objective is to minimize the clinic's closing time, which is closely related to providers' ending times, we found empirically that (3.1f) works in conjunction with (3.1a) to reduce the computational effort during branch and bound. Network and routing models typically use the Miller-Tucker-Zemlin constraints for the

same purpose as (3.1f), but those constraints include a term equivalent to T_{max} to ensure redundancy when necessary (see Miller et al. 1960). Such formulations are known to provide weak LP relaxations, and proved to be ineffective when trying to solve the stochastic version of the IPU scheduling problem.

3.3.3 Room Constraints

In this section, we present our model for the room constraints. Recall that before a patient can be seen by a provider, he is assigned to one of R rooms and remains there until all provider visits are completed. At that point, the room is released and available for the next patient to occupy. When all rooms are in use, arriving patients must wait.

We proposed and tested three methods that equivalently limited the use of rooms to the number available without allowing patients to overlap in the same room. One method may be better than the others, depending on the specific problem. For example, when the number of providers is increased or decreased, the relative performance of the three methods also changes. The most efficient method for our IPU scheduling problem is based on network flow and is presented below. The other two methods are outlined in Appendix B.3.

Network method. The key variables in this approach are T_j^1 and T_j^2, for all $j \in J$, which appear in constraints (3.1k) - (3.1m). Now define a new variable z_{ij} to be 1 if patients i and j use the same room in immediate succession, and 0 otherwise. Let $N = J \cup \{o\}$ be a set of nodes in a network that models patient flow through the clinic, where o is a dummy source/sink node. Between every two nodes in N, we introduce an undirected edge with lower bound 0 and upper bound 1. At the source node, we set the outflow and inflow to be R, and at the patient nodes we set the outflow and inflow to be 1. The patient nodes that receive inflow from the source node correspond to the patients who are the first to use a room. The other flows correspond to the order in which the

patients are assigned to rooms.

Let z_{ij} be the flow from node i to node j, for all $i \neq j \in N$. The constraints for room requirement are as follows.

$$\sum_{j \neq i, j \in J \cup \{o\}} z_{ij} = \begin{cases} R, & i = o \\ 1, & i \in J \end{cases} \tag{3.3a}$$

$$\sum_{i \neq j, i \in J \cup \{o\}} z_{ij} = \begin{cases} R, & j = o \\ 1, & j \in J \end{cases} \tag{3.3b}$$

$$T_j^1 \geq T_i^2 - (1 - z_{ij}) \cdot T_{max}, \quad \forall\, i \neq j \in J \tag{3.3c}$$

$$\sum_{m=1}^{m^k} m \cdot x_{jm}^k \geq \sum_{m=1}^{m^k} m \cdot x_{im}^k + 1 - (1 - z_{ij}) \cdot m^k, \quad \forall\, i \neq j \in J(k), \ \forall\, k \in \{k : n^k = 1\} \tag{3.3d}$$

$$z_{ij} \in \{0, 1\}, \quad \forall\, i, j \in J \cup \{o\} \tag{3.3e}$$

Constraints (3.3a) and (3.3b) specify the outflow and inflow at the nodes, respectively, and together preserve flow balance. Constraints (3.3c) guarantee that a patient's starting time is no earlier than his immediate predecessor's ending time. Constraints (3.3d) are useful cuts, which state that if patient i leaves his room earlier than patient j enters the room, then patient i's position index should be smaller than patient j's position index for any provider who is the only provider of his type. The difference must be at least 1. Constraints (3.3e) define the variables.

3.3.4 Application to Joint Pain IPU

In this section, we adapt the EOSS model (3.1) to the joint pain IPU at the Dell Medical School. Provider types include nurse practitioners, surgeons, physical therapists, nutritionists and care planners. The clinic currently operates with two nurse practitioners and one each of the other

four provider types. As shown in Figure 3.2, after self check-in and rooming, every patient is first seen by a nurse practitioner. Depending on the chief complaint, the patient may be seen by one or more of the next three providers. If the patient requires a consult with the surgeon, this takes place immediately after the nurse practitioner. The physical therapist and nutritionist can be seen in any order. Finally, every patient must meet with the care planner at the end of the visit. After a provider finishes with a patient, the next provider can enter the room immediately but the former provider must complete a small number of administrative tasks (e.g., writing prescriptions) before moving on to her next patient. In the join pain IPU, 7 exam rooms are available for treatment

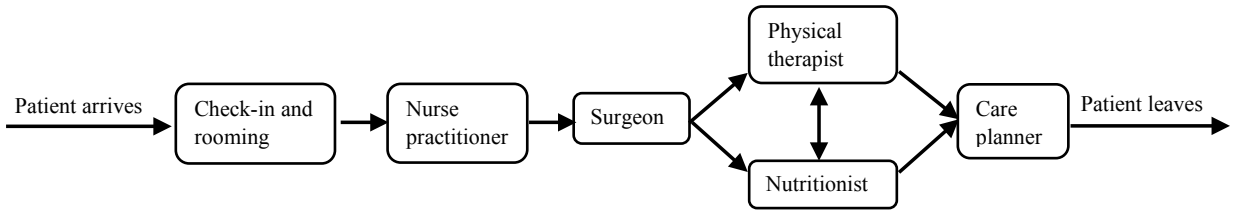


Figure 3.2: Patient paths in joint pain IPU

and consultation. Once assigned to a room, the patient remains there until his visit with the care planner ends and he departs.

3.3.4.1 Discrete-time Arrival

If there are no other constraints on arrival times, then patient j's appointment time will be T_j^1 minus the time for check-in and rooming. In practice, however, clinic appointment times are assigned at fixed intervals rather than continuously throughout the day as the solution to model (3.1) would indicate since t_m^k is a continuous variable. For example, if the clinic opens at 8:00 am and we use a 15-minute interval, then patients can be scheduled at 8:00, 8:15, 8:30, ... Assume that each patient spends s_0 minutes on check-in and rooming. Let τ be the minimum time between scheduled

appointments and let q be the index for arrival time points. Also, let $x_{jq}^{arr1} = 1$ if patient j arrives at the q^{th} time point (multiple of τ) and sees the first nurse practitioner, and 0 otherwise. Let $x_{jq}^{arr2} = 1$ if patient j arrives at the q^{th} time point and sees the second nurse practitioner, and 0 otherwise. We define a new variable n_q^{arr} to represent the schedule template such that $n_q^{arr} = \sum_j (x_{jq}^{arr1} + x_{jq}^{arr2})$ indicates the total number of patients who arrive at time point q. The following constraints are needed for the discrete-time arrival requirement (for convenience, it is assumed that T_{max} is an integral multiple of τ).

$$\sum_{q=0}^{T_{max}/\tau-1} (x_{jq}^{arr1} + x_{jq}^{arr2}) = 1, \qquad \forall\ j \in J \tag{3.4a}$$

$$T_j^1 \geq s_0 + \sum_{q=0}^{T_{max}/\tau-1} (x_{jq}^{arr1} + x_{jq}^{arr2}) \cdot q \cdot \tau, \quad \forall\ j \in J \tag{3.4b}$$

$$n_q^{arr} = \sum_{j \in J} (x_{jq}^{arr1} + x_{jq}^{arr2}), \qquad q = 0, 1, \ldots, T_{max}/\tau - 1 \tag{3.4c}$$

$$x_{jq}^{arr1}, x_{jq}^{arr2} \in \{0,1\},\ n_q^{arr} \in \{0,1,2\}, \quad \forall\ j \in J,\ q = 0, 1, \ldots, T_{max}/\tau - 1 \tag{3.4d}$$

Constraints (3.4a) ensure that each patient arrives at the clinic at one of the T_{max}/τ time points. Constraints (3.4b) guarantee that each patient j is checked in and roomed before being seen by his first provider. Constraints (3.4c) determine the number of patients who arrive at each time point. Constraints (3.4d) define the variables, where for practical purposes the maximum number of patients who are permitted to arrive at any time point is limited to 2. When this bound is relaxed, we found it rare that more than two patients are assigned the same appointment time. Because our ultimate goal is to derive appointment templates that are near-optimal for a large number of scenarios with both stochastic service times and patient pathways, a handful of violations will have a negligible effect on the results.

3.3.4.2 Valid Inequalities – Lower Bounds

The joint pain IPU treats two types or groups of general patients: new and follow-up. New patients usually require longer service times with providers than follow-ups. It is assumed that the ratio of the two patient types is an input parameter. One decision that the model makes is the ordering of the patient types. When a patient calls to schedule a visit, it is known whether he is a new or follow-up patient. Therefore, the arrival time can be set based on one of several rules, such as "all follow-ups at the end of the session." Other information about the patient, such as which providers he will see and their service times, is not known when the appointment is made. That is, the patient routing is determined after the nurse practitioner encounter during which a diagnosis is made.

Since every patient is assumed to spend the same amount of time for check-in and rooming, they see the nurse practitioner in a first-come, first-served order. This allows us to calculate a lower bound on each patient's starting time with the nurse practitioner. Using similar reasoning, if patient i starts no later than patient j, and there are other patients who start no later than patient j but no earlier than patient i, we can also find a lower bound on the time interval between patient i and patient j's starting time with the nurse practitioner.

Specifically, let A_j be the set of patients of the same type as patient j whose visit with the nurse practitioners starts no later than patient j's, excluding j. Let $M(A_j, n)$ be the sum of the n largest service times with the nurse practitioner of the patients who belong to set A_j. This leads to the following proposition which provides a lower bound on the patients' starting times with the nurse practitioner.

Proposition 11 (Separation Proposition)**.** *If patients j_1 and j_2 are of the same type, and patient*

j_1's visit with a nurse practitioner begins no later than patient j_2's, then

$$ST^1_{j_2} - ST^1_{j_1} \geq \Big(\sum_{i \in A_{j_2} \setminus A_{j_1}} s^1_i - M(A_{j_2} \setminus A_{j_1}, n^1 - 1) \Big) / n^1 \tag{3.5a}$$

where n^1 is the number of type 1 providers (nurse practitioners).

Proposition 11 provides a lower bound on the time interval between the starting times of any two patients with a nurse practitioner. Since the first patient always sees the nurse practitioner at $t = s_0$, by applying Proposition 11 to the first patient and any other patient j, we can get a lower bound on any patient j's starting time with a nurse practitioner. Adding such constraints to model (3.1) greatly speeds up the computations because they eliminate many alternative sub-optimal sequences while giving a tighter LP relaxation. These improvements were confirmed during testing.

3.3.5 A Two-Step Method to Solve the Deterministic Problem

The clinic scheduling problem depicted in Figure 3.2 is a combination of an open shop and flexible flow shop problem that turns out to be extremely difficult to solve with a commercial code such as CPLEX for more than 10 patients. To obtain solutions, we developed a two-step method that provides both lower and upper bounds as well as a feasible solution to the original problem.

In Step 1, we remove a subset of the original constraints to create a much easier problem. The relaxed solution provides a lower bound on the objective function in (3.1a) but is rarely, if ever, feasible. In Step 2, we solve a second optimization problem that makes use of the patient sequence found in Step 1. In choosing the constraints to remove in Step 1, we were guided by the speed-up observed after tentatively removing a set of constraints as well as the relative value of the lower bound obtained. For our problem, we found that the best compromise was to remove the following

two sets of constraints.

(a) Room constraints. There were three reasons for this decision. First, removing the room constraints led to only a small decrease in the objective function value. Second, only minimal violations of the constraints were observed, and third, the problem became much easier to solve since many of the binary variables could also be removed.

(b) Nurse practitioner constraints. As the number of providers decreases, the problem gets easier to solve and still provides a lower bound. The decision to omit the nurse practitioners was made for two reasons.

(i) Given that all patients must see a nurse practitioner first, this is the only provider whose waiting time can be taken into account after she is removed from the model. In the original problem, the total delay of patient j attributable to a type k provider consists of two parts: (1) service time s_j^k with the provider, and (2) waiting time when the provider is occupied with prior patients. If we remove the type k provider from the problem without taking into account one or both of these times, the likelihood of getting a strong lower bounds is not very high. The advantage of removing the nurse practitioner rather than any of the other providers is that we are able to connect the starting time of a patient's encounter with the nurse practitioner to the patient's arrival time. For example, if patient j arrives at time point $q-2$, and $s_j^1 + adm^1 - 2\tau > 0$, then any patient who sees the same nurse practitioner as j and arrives at time point q would need to wait for at least $s_j^1 + adm^1 - 2\tau$ minutes before seeing this nurse practitioner. This calculation is myopic and therefore provides a lower bound of the true waiting time.

Let t_j^{wait} denote such a lower bound, and for convenience let $\overline{m} = \max_{j \in J}(s_j^1 + adm^1)/\tau$. The constraints below are needed to determine t_j^{wait}. In each constraint,

patient j's waiting time should be no less than the delay caused by prior patients who see the same nurse practitioner.

$$t_j^{wait} \geq \sum_{i \in J} x_{i,q-m}^{arr1} \cdot (s_i^1 + adm^1) - m \cdot \tau - (1 - x_{j,q}^{arr1}) \cdot T_{max},$$
$$\forall\ j \in J,\ m = 1, \ldots, \overline{m} \tag{3.6a}$$
$$t_j^{wait} \geq \sum_{i \in J} x_{i,q-m}^{arr2} \cdot (s_i^1 + adm^1) - m \cdot \tau - (1 - x_{j,q}^{arr2}) \cdot T_{max},$$
$$\forall\ j \in J,\ m = 1, \ldots, \overline{m} \tag{3.6b}$$

(ii) Because there are two nurse practitioners and every patient must be seen by one of them, the number of binary variables and constraints needed to model this encounter is much greater than for the other providers. Therefore, removing the nurse practitioners greatly reduces the size of an instance and was seen to reduce runtimes by almost an order of magnitude.

Based on the solution from Step 1, we construct a feasible solution to the original problem in Step 2 by adding back the room constraints and solving a modified optimization problem that makes use of patient order. The details follow.

Two-Step Method

Step 1 **(a)** *Preprocessing*. Modify model (3.1) as follows: remove all the variables that have index $k = 1$; remove the nurse practitioner constraints, which are those in model (3.1) for $k = 1$; add constraints (3.6) to model (3.1); subtract t_j^{wait} and s_j^k from the right-hand side of constraints (3.1k) to account for the delay associated with waiting for and being treated by a nurse practitioner after check in.

(b) *Solution*. Set up and solve the relaxed model which consists of the modifications made to model (3.1) in part (a), constraints (3.4), and constraints (3.5a) in Proposition 11.

(c) *Output*. Each patient's appointment time at the clinic. These values can be calculated from x_{jq}^{arr1} and x_{jq}^{arr2}, $\forall\, j \in J, q = 0, 1, \ldots, T_{max}/\tau - 1$.

Step 2 **(a)** *Preprocessing*. Order the patients based on their arrival time in the solution found in Step 1. Each patient has a rank order.

(b) *Model modifications*. Construct a new model, which includes model (3.1), constraints (3.3), constraints (3.4), and constraints (3.5a) in Proposition 11. Also add the following constraints: if patient j's rank order is two or more greater than patient i's in the solution found in Step 1, then $\sum_{q=0}^{T_{max}/\tau-1}(x_{jq}^{arr1} + x_{jq}^{arr2}) \cdot q \cdot \tau \leq \sum_{q=0}^{T_{max}/\tau-1}(x_{iq}^{arr1} + x_{iq}^{arr2}) \cdot q \cdot \tau$. Accordingly, j will arrive no earlier than patient i in the new solution.

(c) *Solution*. Set up and solve the model resulting from the modifications found in part (b).

(d) *Output*. Each patient's appointment time and the schedule template for the clinic.

3.4 Stochastic Model

The deterministic EOSS model formulated in Section 3.3 can be used to solve an instance of the daily appointment scheduling problem but it falls short in accounting for the stochastic elements in the system. Our real goal is to develop an appointment template that is robust in the face of probabilistic service times and patient flows. A priori uncertainty in routing is the norm when patients are to be seen by multiple providers in a single visit. In fact, it is the rule rather than the

exception in many clinical settings, since the personalized plan of care is made after the patient has been initially interviewed and examined to determine the severity of his condition. Therefore, it is not possible to accurately predict which providers he will need to see during the visit.

Based on our deterministic model, we have developed a two-stage integer stochastic programming model in which the patient mix along with service times, provider sets and pathways are random variables. The objective of the model is to minimize a weighted combination of expected clinic closing time and patient waiting time over a wide range of scenarios. In the accompanying analysis it is assumed that the no-show rate is zero and that all patients arrive at their scheduled time.

3.4.1 Stochastic Problem

In our IPU scheduling problem, the likelihood that a patient sees a particular provider for a specific amount of time is determined by probability distributions obtained from the Dell Medical School Department of Surgery. For lower extremity joint pain, new and follow-up patients are further divided into six sub-types: (new) mild osteoarthritic, moderate osteoarthritic, severe osteoarthritic, operative, follow-up non-operative, and follow-up operative. Given their proportional mix and their associated probability distributions for provider sets and service times, it is possible to generate scenarios using Monte Carlo sampling. Our original intent was to generate half-day scenarios (4.5-hour clinical sessions) and then try to solve the corresponding two-stage stochastic program to determine the optimal appointment template. We found, however, that as the number of scenarios grew it was increasingly difficult to find solutions, so various alternatives to tackling the full problem were investigated. In the simplest case, we find a template and corresponding patient flow for each scenario separately by solving the corresponding deterministic EOSS model. The average clinic closing time and patient waiting time are then calculated over the different scenarios to get a

lower bound on long-run clinic performance. This is called the *wait-and-see* (WS) solution.

In the first stage of the two-stage model, a single appointment template is determined without knowing the patient mix, provider sets, pathways, and service times. In the second stage, this information is revealed for each scenario. To formulate the problem, denote the patients' provider sets and service times by $\tilde{K}$ and $\tilde{s}$, respectively. Assuming for the moment that the appointment template is known, we can then find the optimal arrival times, room occupancy times, and provider start and end times with their patients for each scenario. That is, we can find the optimal values of the the second stage variables, which we denote by $\hat{x} \equiv \{x, y, z, x^{arr}, t, ST, T^1, T^2\}$. These values specify each patient's arrival time and schedule with his providers. Letting $n^{arr} \equiv (n_0^{arr}, n_1^{arr}, \ldots, n_{T_{max}/\tau}^{arr})$ be the arrivals at time point q, the two-stage stochastic program, also known as the *recourse problem* (RP), is

$$\min_{n^{arr}} \; E_{\tilde{s},\tilde{K}}\Big[f(n^{arr}, \tilde{s}, \tilde{K})\Big] \tag{3.7a}$$

$$\text{s.t. Constraints (3.1b)} - \text{(3.1o)}, \text{ (3.3a)} - \text{(3.3e) and (3.4a)} - \text{(3.4d)} \tag{3.7b}$$

where $E_{\tilde{s},\tilde{K}}$ denotes the expectation with respect to the random variables $\tilde{s}$ and $\tilde{K}$, and $f(n^{arr}, \tilde{s}, \tilde{K})$ is defined as

$$f(n^{arr}, \tilde{s}, \tilde{K}) = \min_{\hat{x}} \; \alpha_1 \cdot T + \alpha_2 \cdot \sum_{j \in J} (T_j^2 - T_j^1)$$

The function $f(\cdot)$ represents the second stage problem. Conceptually, after the appointment template n^{arr} is fixed in the first stage, all uncertainty is resolved and optimal schedules can be determined in the second stage for each patient in each scenario. For a fixed template, the individual scenario instances can be solved separately (we solve each scenario using our deterministic model presented in Section 3.3) and their objective values averaged to get an approximation of the objective function value in (3.7a). This approach is called sample average approximation (e.g., see

Kleywegt et al. 2002).

3.4.2 Solving the Stochastic Model

When the number of scenarios is finite, the two-stage stochastic program is typically approached by creating a deterministic equivalent one-stage, mixed-integer program. In the reformulation, the second-stage constraints and variables are indexed by scenario and the expected value in (3.7a) is replaced with the average of the second-stage objective functions (e.g., see Bard et al. 2007; Engell et al. 2004). However, such an approach does not always work well because the computational burden increases dramatically as the number of scenarios increases. This was the situation that we faced after enumerating only a few scenarios.

The first alternative that we investigated involved replacing the random parameters with their expected values to obtain a deterministic formulation known at the expected value (EV) problem. For IPUs, however, the likelihood of a patient seeing a particular provider follows a probability distribution, so taking the expectation of the patient's provider set would lead to fractional visits. To deal with this situation we conducted a Monte Carlo simulation by sampling each patient's provider set to generate different scenarios. In each scenario, we used the expected service times and expected number of patients of each of the six types (rounded to the nearest integer). The optimization problem for each scenario is solved using our deterministic model in Section 3.3. After finding the solution for each scenario, we average the numbers of patients who arrive at each time point in all scenarios to get the expected value solution. The EV problem can be stated as follows:

$$EV = \min_{n^{arr}} \; E_{\tilde{K}}\left[f(n^{arr}, \hat{x}, E[\tilde{s}], \tilde{K}) \right] \tag{3.8a}$$

where the optimal objective function value is denoted by EV and the value of the template variables is denoted by n_{EV}^{arr}. We are also interested in the solution of the following three problems which are used to evaluate the quality of the EV solution and to calculate upper and lower bounds on the optimal solution.

$$RP = \min_{n^{arr}} \; E_{\tilde{s},\tilde{K}}\left[f(n^{arr},\hat{x},\tilde{s},\tilde{K})\right] \tag{3.9a}$$

$$EEV = E_{\tilde{s},\tilde{K}}\left[f(n_{EV}^{arr},\hat{x},\tilde{s},\tilde{K})\right] \tag{3.9b}$$

$$WS = E_{\tilde{s},\tilde{K}}\left[\min_{n^{arr}} f(n^{arr},\hat{x},\tilde{s},\tilde{K})\right] \tag{3.9c}$$

RP represents the two-stage stochastic program given by model (3.7), and as mentioned, is not solvable; hence the need for bounds. To measure the quality of the EV solution, we fix the template in RP to n_{EV}^{arr} and solve the resulting second-stage problems separately. Averaging their objective function values gives what is called the expected cost of the EV solution, which is denoted by EEV. The EEV value is an upper bound on *RP* and WS is a lower bound (see Birge and Louveaux 2011). Thus we have the following relationships.

$$WS \leq RP \leq EEV$$

The optimality gap associated with EV is the gap between EEV and *RP*, which derives from the loss of stochasticity in the EV problem. The gap between WS and *RP* results from the loss of perfect information. Neither of these gaps are possible to obtain in our case, though, because we are not able to find *RP* for realistic instances. Therefore, we turn to the gap between WS and EEV to evaluate the quality of the EV solution. Since the WS and EEV problems are solved using the two-step method, we use the gap between the step-one value obtained from the WS problem, and

the step-two value obtained from the EEV problem to evaluate performance.

3.5 Computational Results

All models were implemented in C++ using IBM's Concert Technology library and solved with CPLEX 12.7. The experiments were performed on a Linux workstation with 4 Intel(R) Core(TM) i7-4790 CPU, 8 3.60GHz processors and 16 GB memory running Ubuntu 16.04. All problem instances discussed in this section were solved optimally using CPLEX's default setting. In constraints (3.1h) and (3.1m), the value of ε was set to 1.2.

3.5.1 Data and Scenarios

In the analysis, we consider half-day sessions consisting of a fixed number of patients. Arrivals are scheduled by the models at multiples of 15-minute intervals beginning at 8:00 a.m. The total time allocated for check in and rooming is 8.3 minutes per patient. The IPU operates with two nurse practitioners and one each of the other provider types. The total number of rooms is 7. Table 3.1 gives the patient mix and the probability that a particular patient type will be seen by each of the providers. The first encounter for all patients is with a nurse practitioner and the last is with the care planner, both with probability 1, so these providers are omitted from the table. As mentioned, the new patients are divided into four groups and the follow-ups into two groups. The ratio between the new and follow-up patients is 3:1.

We model the probabilities for a certain type of patient seeing each of the different providers as independent. This reflects the fact that we do not know a given patient's path a priori. Whether a patient sees a certain provider is determined after the patient arrives at the clinic and is examined by the nurse practitioner. Under such circumstances, it is common to take a population-level view and use independently sampled probabilities (see, e.g., Lahiri and Seidmann 2012, White et al.

2011, Dobson et al. 2013, and Saghafian et al. 2014). Service time distributions are enumerated in Table 3.2.

Table 3.1: Patient probabilities for visits with providers

Patient type	Patient mix	Surgeon	Physical therapist	Nutritionist
New mild osteoarthritis	0.330	0.25	0.5	0.4
New moderate osteoarthritis	0.3225	0.5	0.5	0.4
New severe osteoarthritis	0.05625	0.9	0.7	0.4
New operative path	0.04125	1	0.9	0.4
Follow-up non-operative path	0.1875	0.3925	0.4875	0.378
Follow-up operative path	0.0625	1	0.5	0

Table 3.2: Service time probability distributions (minutes)

Patient type	Nurse practitioner	Surgeon	Physical therapist	Nutritionist	Care planner
New mild osteoarthritis	Tri(15,20,30)	Tri(7,10,20)	Tri(10,15,25)	Tri(10,15,25)	Tri (5,10,20)
New moderate osteoarthritis	Tri(15,20,30)	Tri(7,10,20)	Tri(10,15,25)	Tri(10,15,25)	Tri (5,10,20)
New severe osteoarthritis	Tri(15,20,30)	Tri(7,10,20)	Tri(10,15,25)	Tri(10,15,25)	Tri (5,10,20)
New operative path	Tri(15,20,30)	Tri(7,10,20)	Tri(10,15,25)	Tri(10,15,25)	Tri (5,10,20)
Follow-up non-operative path	Tri(7,12,17)	Tri(4.3,5.4,10.8)	Tri(10.8,16.2,21.6)	Tri(6,8,12)	Tri (5,10,20)
Follow-up operative path	Tri(7,12,17)	Tri(3.024,4.32,8.64)	Tri(6.48,8.64,12.96)	0	Tri (5,10,20)

The implied pathways and probability distributions in Tables 3.1 and 3.2 are based on estimates provided by the director of the lower extremity joint pain IPU in the musculoskeletal area at the Dell Medical School (DMS) (fourth author on this paper) and other providers from the DMS Department of Surgery who had experience with the same patient population at other clinics prior to the formation of the joint pain IPU. The six patient types (pathways) identified in the two tables represent a common characterization of patients seeking treatment for joint pain. This level of detail allowed the clinical team to estimate the probabilities associated with the resources required to provide care to each type of patient. At the highest level, patients are generally classified as new or follow-up. Clinically speaking, there are only two types of follow-up patients. Those that follow up after surgery and have a certain type of pathway resulting in a fairly short and predictable visit, versus a non-operative follow-up visit, which is similar across disease severity and somewhat

longer than a postoperative visit. In rare cases, some patients may benefit from supplementary services such as psychiatry, social work and behavioral health. However, having dedicated providers to cover these services could not be justified financially so they were not included in the design of the IPU.

In the absence of historical data, anecdotal evidence suggests that the time to undergo medical procedures in an outpatient setting can be modeled using minimum, maximum and modal times (e.g., see Swisher et al. 2001). These three parameters, solicited from the aforementioned providers, lead directly to a triangular distribution, which we use for service times. As an aside, when the clinic opened in the fall of 2017, the staff was able to collect data on provider service times and patient mix. This led to a few adjustments in the probabilistic data in Tables 3.1 and 3.2, but for the most part, the original estimates turned out to be highly accurate.

In our evaluation of the two-step method in Section 3.5.2, parameter values and provider sets for each patient are generated independently. First we determine which type of patient is being considered by sampling from the patient mix distribution in Table 3.1. Although the total number of patients is fixed in each scenario, the ratio of new to follow-ups changes from one realization to the next. After each patient's group is determined, we generate his provider set according to the probabilities in Table 3.1, and service times from the triangular distributions in Table 3.2. The same generated data sets are used for the EEV and WS problems.

When deriving the EV template defined in Section 3.5.3.1, rather than sampling from the patient mix distribution, the number of new and follow-up patients was set to their approximate expected values. For each patient type, the provider set was generated according to the probabilities in Table 3.1, while the expected service time with each provider was taken as the weighted sum (the weight is the patient mix fraction) of the mean service time. For example, the expected service time of a follow-up patient with the surgeon is the weighted sum of the mean of the bottom two

triangle distributions under the column 'Surgeon' in Table 3.2. Lastly our models reflect whether a patient is new or making a follow-up appointment at the time of booking. In practice, this is all the information that is available to the scheduling clerk.

3.5.2 Two-Step Method

In the first set of experiments, our goal was to evaluate the quality of the solutions obtained with the two-step method presented in Section 3.3.5 for solving the deterministic model. We began by randomly generating 200 instances (scenarios) with 10 patients each and then applying the algorithm. The number of patients in each instance was determined by sampling from a multinomial distribution with probabilities $\{0.3, 0.2, 0.1, 0.1, 0.2, 0.1\}$, which approximates the patient mix in Table 3.1. Similarly, the provider set for each type of patient was sampled using the probabilities in Table 3.1 while the service times were sampled from the triangular distributions in Table 3.2. Recall that Step 1 provides a lower bound and Step 2 provides an upper bound on the objective function in (3.1a). Performance was measured by the percentage deviation from the optimum obtained by solving model (3.1) as modified to represent the joint pain IPU. We only considered instances with 10 patients in this part of the analysis because it was not possible to reliably solve larger instances with CPLEX. Note that after 200 instances, the output statistics discussed below were unchanged to two decimal places, indicating that there was no further need for additional sampling. In all, 16,096 seconds were required to find the exact optima for the 200 instances compared to 1935 seconds when using the two-step method to find the bounds.

For each scenario, we calculated the gap between the Step 2 objective function value and the Step 1 value (GAP 2-1), the gap between the Step 2 value and optimal value (GAP 2), and the gap between the optimal value and the Step 1 value (GAP 1). The differences were then converted to percentages and averaged over the 200 scenarios. The results are summarized in Table 3.3.

Table 3.3: Optimality gap for the two-step method with 10 patients

Statistics	GAP 1	GAP 2	GAP 2-1
Mean	2.73%	0.97%	3.69%
HW[1]	0.39%	0.21%	0.49%

[1] Half width of a 95% confidence interval

From the table we see that the average gap between the bounds found in Steps 1 and 2 is 3.69%, an indication of the strength of the heuristic. Additional evidence of its strength can be seen by examining the percent difference between the upper bound and the optimal solution (GAP 2), which is only 0.97% on average. Moreover, the optimal solution is much closer to the Step 2 solution than the Step 1 solutions because GAP 2 is a third the size of GAP 1. Taken together, these results support the use of the two-step method to derive appointment schedules under more realistic scenarios.

To check the sensitivity of the performance of the two-step method, we repeated the above process for cases with 7, 8 and 9 patients. The results are reported in Table 3.4. The optimality gap decreased slightly as the number of patients decreased but remained stable. In our testing with 14 patients in the remaining sections, the gap was always less than 5%.

Table 3.4: GAP 2-1 for the two-step method with different numbers of patients

Number of patients	7	8	9	10
Mean	2.38%	3.14%	3.61%	3.69%
HW[1]	0.39%	0.45%	0.44%	0.49%

[1] Half width of a 95% confidence interval.

3.5.3 Finding Robust Templates

Our primary goal is to derive a single appointment template whose implementation will assure clinic durations of less than 4.5 hours and patient visit times not exceeding 1.5 hours, on average.

The recourse problem was designed to achieve this goal but the computational difficulties we encountered when trying to solve it led to our reliance on the two-step method. The best we can do with this heuristic, however, is to solve a deterministic version of model (3.1). The approach we take to circumvent this limitation is described below. For the remaining analysis, we work with 14 patients, which is the number that the joint pain IPU would like to schedule each half-day session.

3.5.3.1 Generating EV Templates

Ordinarily, only a single EV template exists, which would be derived by replacing all random parameters in the IPU model with their expected values and then solving. This was not possible for our problem because the expected number of providers that sees a patient is a random variable whose expected value is fractional. As mentioned, Monte Carlo sampling was used to skirt this issue. The first step was to generate a representative number of scenarios by using the data in Tables 3.1 to obtain the provider set for each patient. As an integral approximation to the patient mix, we assumed that each scenario consisted of 11 new patients and 3 follow-ups. For the former group, the number of patients of each type was fixed at 4, 4, 2 and 1. For the latter group, the number of patients was fixed at 2 and 1. We then used the two-step method to find feasible schedules and their corresponding templates n^{arr}, where n^{arr} is a vector that specifies the number of patients who arrive at each 15-minute time point.

To derive a single appointment template, we began by averaging the number of patients who arrive at each time point over all scenarios. Again we found that the output statistics became stable after 200 scenarios so we terminated the generation process at that point. The total time required to solve the 200 instances was 114 minutes. Figure 3.3 depicts the results after averaging. The horizontal axis indicates the time points and the vertical axis identifies the average number of patients who are scheduled to arrive at the start of each 15-minute interval.

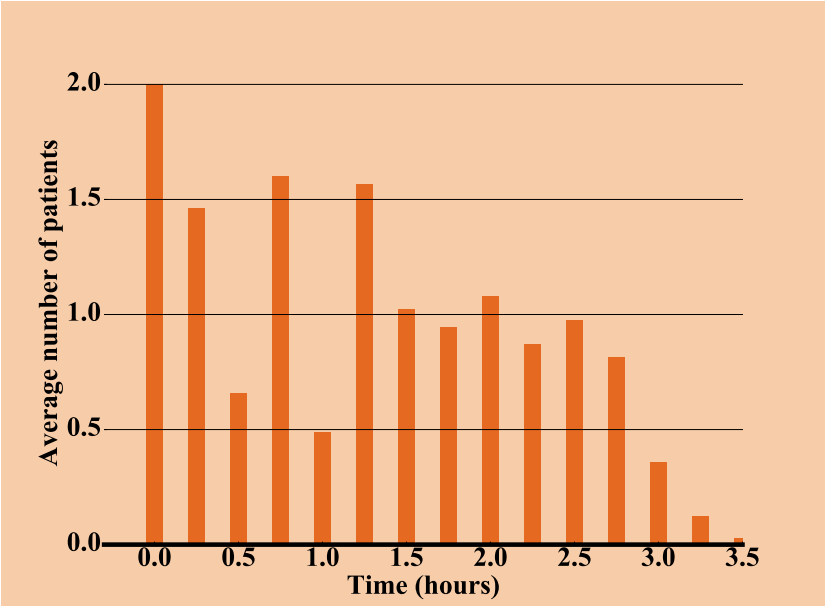


Figure 3.3: Average number of patients scheduled to arrive at each time point

Rounding strategies. As expected, the height of the bars in the figure are fractional, but to be implementable the number of patients must be 0, 1 or 2 at each time point, as in the individual solutions. To achieve integrality, a rounding strategy is necessary. The approach we take is based on the observation that the number of patients who arrive earlier in the session affect the statistics of patients who arrive later. Accordingly, the procedure we adopt is to round fractions (up or down), fix the number of patients at one point at a time starting at zero, and sequentially moving forward in 15-minute increments until closing time is reached. At each time point t, the number of patients who have arrived previously is fixed by rounding. We then round the fractional number at t and repeat the procedure at $t+1$.

In particular, after fixing the number of patients who arrive at t, we re-solve the reduced EV problem with the remaining patients, average the results from the newly derived 200 templates, and then round the value at $t+1$. For example, at $t=0$, we see in Figure 3.3 that the average

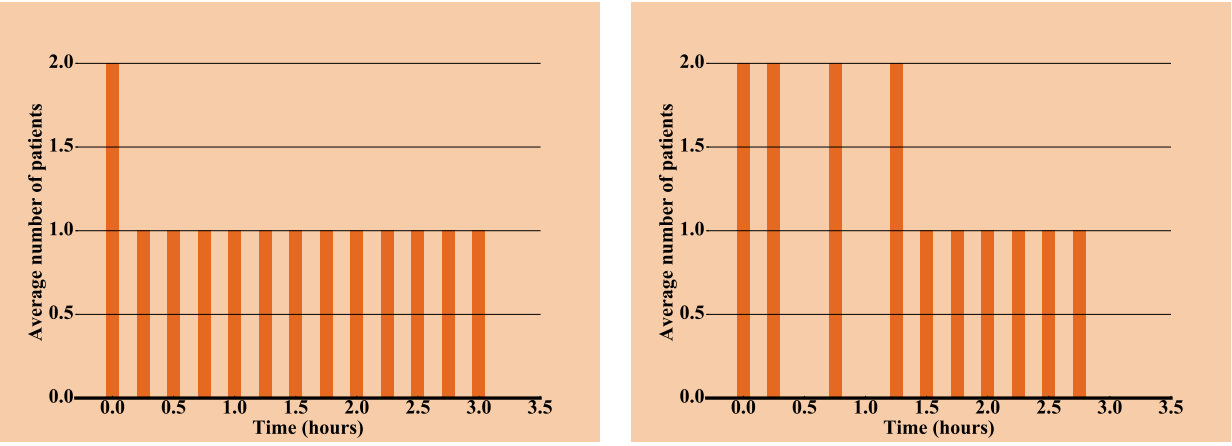
number of patients is very close to 2 so we fix the number of patients who are scheduled to arrive at $t = 0$ to 2; that is, we set $n_0^{arr} = 2$. We then re-solve the EV problem and take the average of the 200 templates just derived. The corresponding figure is almost identical to Figure 3.3, so with $n_0^{arr} = 2$, we fix n_1^{arr} to be either 1 or 2 depending on the rounding strategy (to follow). After fixing n_1^{arr}, we re-solve the EV problem and move on to n_2^{arr}, and so on.

The number of possible templates increases exponentially with the number of time points for arbitrary rounding. We considered two strategies to generate two templates. In the first strategy we always round up at t unless the fraction is zero or within a small range of an integer value. Based on empirical testing, we chose the cutoff to be 0.2. If the average number of patients is less than 0.2, we round it to 0; if it is between 0.2 and 1.2, we round it to 1; if it exceeds 1.2 but is less than 2, we round it to 2. Without a cutoff we found that the resulting schedules were too aggressive in that they emphasized earlier appointment times, which led to significantly longer patient waiting times.

In the second strategy, we always round down at each time point, unless the fraction is within the cutoff range. Based on empirical testing, we again chose the cutoff to be 0.2. If the average number of patients is less than 0.8, we round it to 0; if it is between 0.8 and 1.8, we round it to 1; if it exceeds 1.8 but is less than 2, we round it to 2.

The template produced by the first strategy is more aggressive than the second but rounding down does not always avoid long waits and extended clinic hours. Figure 3.4 shows the less aggressive EV template [panel (a)] and the more aggressive EV template [panel (b)]. Each panel indicates the number of patients scheduled to arrive at each time point. Note that during construction, the last patient in the less aggressive template actually arrives at $t = 3.25$. For practical reasons, though, we modified the template slightly to avoid a gap at $t = 3.0$ and to conform with what is called the 2BEG schedule in the literature (Cayirli and Veral 2003). Testing showed negligible differences

between results produced by the less aggressive template and 2BEG.



(a) Less aggressive EV template (b) More aggressive EV template

Figure 3.4: Templates derived from the EV solution

Comparison of strategies. To visualize the difference between the more aggressive and less aggressive templates, we generated the cumulative number of patients who arrive at the clinic up to each time point t. Of course, the total number of arrivals for the less aggressive template is no greater than that for the more aggressive template at any t. Figure 3.5 plots the results as a function of time for both templates. Any other template that is constructed from a combination of the less aggressive and more aggressive strategies would be bounded by these two curves. Comparing the cumulative number of patients for the two schedules at any time t shows that the difference is small. In fact, the two plots in Figure 3.5 indicate that the difference at any time t is either 0 or 1.

Additional templates. In addition to the two templates derived above, we also evaluated a third from the literature and a fourth based on a variation of the more aggressive template in Figure 3.4(b). Each of the four templates is formally defined below and consists of the number of patients who arrive between $t = 0$ and $t = 3$ (i.e., between 8 am and 11 am), followed by its name and description.

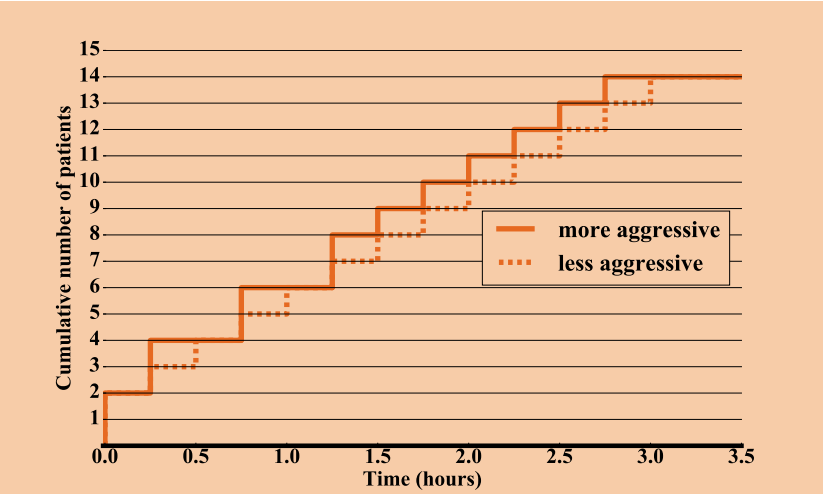


Figure 3.5: Cumulative number of patient arrivals over a half-day session

- 2-1-1-1-1-1-1-1-1-1-1-1-1: 2BEG. Assigns two patients at the beginning of the session and then one at each point thereafter. It was first proposed and studied in Bailey (1952), and turns out to be the less aggressive EV template that we derived.

- 2-0-2-0-2-0-2-0-2-0-2-0-2: VBFI-1. VBFI stands for 'variable block/fixed interval,' which means that a different number of patients can be assigned at each time point as long as they are separated by the same fixed interval (see Wijewickrama 2006). Here, two patients are scheduled to arrive every half hour. This is less aggressive than 2BEG, which can be transformed into VBF-1 by moving one patient at every other time point to the next time point starting at $t = 0.25$. In our experience, VBFI-1 is commonly used in practice.

- 2-2-0-2-0-2-1-1-1-1-1-1-0: EV-RU. This is the more aggressive template shown in Figure 3.4(b), where RU stands for 'round up.'

- 2-2-0-2-0-2-2-0-2-0-2-0-0: VBFI-2. This template is based on EV-RU but is more aggressive. If we move the patients at $t = 1.75$, 2.25 and 2.75 in the EV-RU template one interval

earlier, we can get VBFI-2. Including this template in the study will tell us whether a significant improvement results by making the EV-RU template more aggressive.

3.5.3.2 Results for Candidate Templates

To compare the quality of the solutions resulting from the use of each of the four templates, we randomly generated additional scenarios by sampling from the distributions in Tables 3.1 and 3.2 to obtain provider sets and service times, respectively, for each patient. The output statistics became stable after 800 scenarios so we stopped at that point. The number of new and follow-up patients are also sampled although the total number was fixed at 14. To gauge performance, we averaged the objective function values and other metrics over all 800 scenarios for each template. The results are highlighted in Tables 3.5 and 3.6 along with the results for the WS problem using the same data. The columns in the tables are arranged from the least aggressive to the most aggressive template. For each template, computation times for all 800 scenarios ranged from a total of 8 to 19 hours.

The first two rows in Table 3.5 report the Step 1 and Step 2 closing times. The remaining rows give the Step 2 flow time statistics for all patients, and then for new patients and follow-ups separately. The last row reports the fraction of cases in which the closing time exceeded 4.5 hours. Table 3.6 shows provider and room utilizations. Because the model includes constraints (3.1m), which restrict the total time a patient can spend in the clinic to a given maximum, a handful of instances turned out to be infeasible. For the EV-RU template, 6 out of 800 were infeasible and for the VBFI-2 template, 10 out of 800 were infeasible.

Table 3.5: Results for different appointment templates

	WS		VBFI-1		2BEG		EV-RU		VBFI-2	
Metrics	Mean[1]	HW[2]	Mean	HW	Mean	HW	Mean	HW	Mean	HW
Step 1 closing time	4.097	0.028	4.438	0.021	4.343	0.022	4.226	0.023	4.205	0.024
Step 2 closing time	4.295	0.024	4.444	0.020	4.360	0.022	4.289	0.023	4.289	0.023
Feasible rate	800/800		800/800		800/800		794/800		790/800	
Waiting time	0.296	0.006	0.281	0.008	0.300	0.009	0.371	0.009	0.414	0.009
Service time	1.080	0.005	1.080	0.005	1.080	0.005	1.079	0.005	1.079	0.005
Time in clinic	1.376	0.010	1.361	0.012	1.381	0.012	1.450	0.013	1.493	0.013
Waiting time (new)	0.310	0.007	0.284	0.009	0.307	0.009	0.381	0.010	0.424	0.010
Service time (new)	1.162	0.005	1.162	0.005	1.162	0.005	1.161	0.006	1.161	0.006
Time in clinic (new)	1.472	0.011	1.446	0.012	1.469	0.013	1.542	0.013	1.585	0.013
Waiting time (follow-up)	0.242	0.010	0.258	0.010	0.268	0.011	0.319	0.012	0.360	0.012
Service time (follow-up)	0.779	0.007	0.779	0.007	0.779	0.007	0.778	0.007	0.779	0.007
Time in clinic (follow-up)	1.020	0.013	1.037	0.013	1.047	0.014	1.097	0.014	1.138	0.014
Fraction above closing time	NA	NA	0.351	0.033	0.263	0.031	0.217	0.029	0.219	0.029

[1] All times in hours; the statistics are all Step 2 results except for the Step 1 closing time.

[2] Half width of a 95% confidence interval.

Table 3.6: Resource utilization

	WS		VBFI-1		2BEG		EV-RU		VBFI-2	
Metrics	Mean	HW	Mean	HW	Mean	HW	Mean	HW	Mean	HW
Nurse practitioner 1	0.731	0.004	0.704	0.004	0.724	0.004	0.732	0.004	0.726	0.004
Nurse practitioner 2	0.712	0.004	0.688	0.004	0.696	0.005	0.712	0.004	0.717	0.004
Surgeon	0.485	0.009	0.468	0.008	0.478	0.008	0.486	0.009	0.486	0.009
Physical therapist	0.686	0.010	0.663	0.010	0.676	0.010	0.684	0.010	0.683	0.010
Nutritionist	0.444	0.011	0.429	0.010	0.437	0.011	0.445	0.011	0.445	0.011
Care planner	0.686	0.005	0.662	0.004	0.675	0.004	0.687	0.005	0.687	0.005
Room	0.549	0.003	0.541	0.003	0.548	0.003	0.569	0.003	0.575	0.003

Theoretically, the Step 2 closing time obtained from VBFI-2 should be no later than the closing time provided by EV-RU for two reasons: (i) VBFI-2 is more aggressive than EV-RU, and (ii) more infeasible cases are discarded when VBFI-2 is used which should bring down the average closing time. This follows because late clinic closing times are a result of long patient waiting times, which produce infeasible instances. Nevertheless, the two templates have virtually identical Step 2 closing times so neither reason was seen to have a noticeable impact on clinic performance. This suggests that the EV-RU template is sufficiently aggressive and that moving to the more aggressive VBFI-2 template will not provide any benefit. This also suggests that there is no bias in the results after discarding the infeasible cases.

Clinic closing time. The first observation from the statistics in Table 3.5 is that the difference between the Step 1 and Step 2 closing times is less than 2% for all four templates. Although the Step 2 closing time in each case is not necessarily optimal, given that the two-step method was used for the computations, the size of the gap indicates that it should be a very good approximation. One way to evaluate the four sets of results is to compare the mean and half width of a 95% confidence interval of clinic closing time of the Step 2 solution. For example, the Step 2 results imply that the 2BEG 95% confidence interval extends from 4.338 to 4.382, while the range of the average clinic closing time for EV-RU is from 4.266 to 4.312. Because the two confidence intervals do not overlap, we can conclude that the average closing time obtained from the EV-RU template is significantly smaller than the value associated with the 2BEG template.

Another way to compare the closing time for different templates is to check the Step 1 and Step 2 solutions. For example, the lower bound on closing time for 2BEG obtained at Step 1 is 4.343, which is greater than the Step 2 closing time of EV-RU. As such, the true value of closing time for 2BEG should also be greater than the true value of closing time for EV-RU. By implication, using the EV-RU template should yield lower clinic closing times than the 2BEG template. For the

WS problem, its optimal clinic closing time should be no greater than the closing time obtained from any template. As can be seen in Table 3.5, however, the average WS Step 2 closing time is 4.295, which is greater than 4.289, the average closing time obtained from the EV-RU and VBFI-2 templates. This result is possible because the two-step method only provides feasible solutions. As it turns out, many of the WS solutions are suboptimal.

A second observation about the statistics in Table 3.5 is that as the templates get more aggressive, the clinic closing times decrease; see Figure 3.6. This follows because patients generally arrive earlier when the more aggressive templates are used, and are seen earlier by their providers. Hence, they are more likely to finish their visit sooner. Because the same 800 scenarios were used in all the computations, the service times are the same across all templates, so the comparative closing time results should not be affected by those values. The statistics in Table 3.5 confirm that the average service time for a visit is nearly identical for all templates as well as for the WS problem.

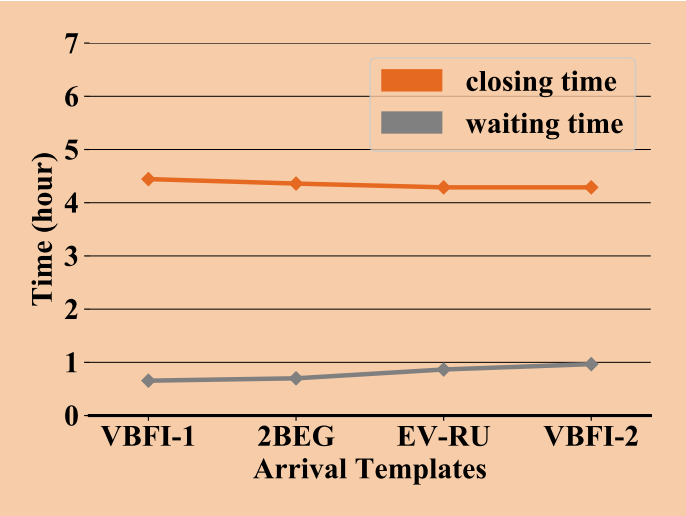


Figure 3.6: Comparison of four arrival templates

Waiting times and time in clinic. The average waiting time and average total time in the clinic

increase as the template becomes more aggressive. Again, more patients arriving earlier makes it more likely that they will face longer queues in front of their providers. This is true for all patients taken as a whole, for new patients, as well as for follow-ups. For example, the waiting time increases from 0.281 to 0.3 to 0.371 to 0.414 hours as the template gets more aggressive. As might be expected, the average waiting time for the WS problem is relatively small even though its average closing time is also small. This follows because a separate template is derived for each scenario allowing patient arrivals to better match provider availability.

As the appointment template becomes more aggressive, the waiting time and closing time move in opposite directions, as can be seen in Figure 3.6. Based on their relative importance, the clinic director can choose the template that achieves the best balance. For example, if preference is given to the closing time, then the EV-RU template may be a good candidate because its closing time is 4.289, which is measurably less than the corresponding value of 4.36 for 2BEG and 4.444 for VBFI-1, a 1.6% (4.3 minutes) and 3.5% (9.3 minutes) reduction, respectively. Moreover, patient waiting times resulting from the EV-RU template increase by 4.3 minutes and 5.4 minutes over 2BEG and VBFI-1, respectively.

The EV-RU template appears to be a good compromise with respect to the primary metrics. If we make it more aggressive by transforming it into the VBFI-2 template, the clinic closing time remains about the same but the patient waiting times increase significantly. Nevertheless, if the waiting time is relatively more important than the closing time, then the 2BEG template may be a good choice because its average waiting time of 0.3 hours is somewhat less than the corresponding values of 0.371 for the EV-RU and 0.414 for VBFI-2 templates. In practice, it is not desirable to choose a template less aggressive than 2BEG such as VBFI-1. The reduction in waiting time provided by the latter is only 1.14 minutes on average, while the average jump in closing time is 5.04 minutes.

Fraction above target closing time. From Table 3.5 we see that the fraction of scenarios in which the clinic closing time exceeds the target of 4.5 hours decreases for the first three templates as they get more aggressive. For VBFI-1 the percentage is 35.1, while for EV-RU the percentage drops to 21.7. There is almost no difference between EV-RU and VBFI-2, which gives further evidence that VBFI-2 does not improve clinic performance even though it is more aggressive than EV-RU.

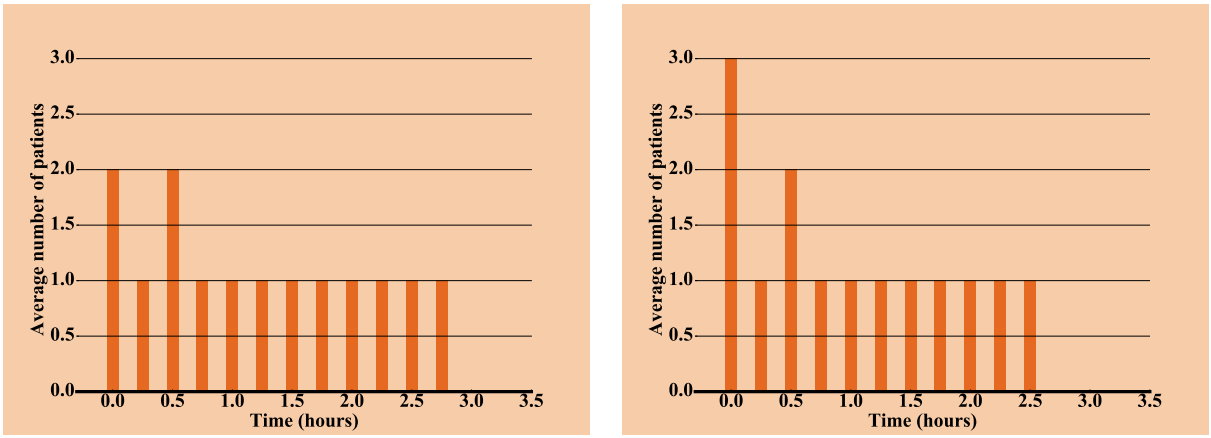
Utilization. Table 3.6 reports the utilization for the six individual providers and the seven rooms. While there are some statistically significant differences between the templates for each provider type, they are negligible in practice. The contrast in room utilization is a bit sharper but still negligible. Note that the values in the table are based on the time the first patient arrives and the last patient leaves. At first glance, the statistics may be somewhat misleading because it takes over an hour for the clinic to fill up and roughly the same amount of time for it to empty out. While waiting times average up to 25 minutes, for example, room utilization is less than 60% on average. This supposed contradiction, can be explained by the transient effects at the beginning and end of the session.

3.5.3.3 Different Resource Levels

To determine the potential value of increasing or decreasing resource levels, we investigated two possibilities. In particular, nurse practitioners and rooms are two resources that afford some leeway in clinic design. Preliminary testing suggested that decreasing or increasing the number of rooms by one barely affected system performance, while increasing the number of nurse practitioners by one had a noticeable impact. Consequently, in this section we only present results for 3 nurse practitioners.

In the analysis, we followed the same procedure outlined in Section 3.5.3.1 using the same data for the patient mix and service time distributions. The two templates shown in Figure 3.7 parallel

those in Figure 3.4. The VBFI-3 template is the less aggressive of the two and VBFI-4 is the more aggressive. Our previous results for these templates still hold. For example, the VBFI-3 template provides better outcomes if the patient waiting time has more weight than the clinic closing time, and the VBFI-4 template is better if clinic closing time is the more important metric.



(a) VBFI-3 (b) VBFI-4

Figure 3.7: Two templates for the case with 3 nurse practitioners. VBFI-3 is the less aggressive EV template, and VBFI-4 is the more aggressive EV template.

It is more interesting, though, to compare the system with 2 and 3 nurse practitioners. The statistical results for these new templates are highlighted in Tables 3.7 and 3.8. A comparison with the statistics in Tables 3.5 and 3.7 indicates that adding one nurse practitioner significantly reduces both the clinic closing time and the patient waiting time. For the more aggressive templates, for example, the clinic closing time decreases from 4.289 to 4.005 hours (6.6%), and the patient waiting time decreases from 0.371 to 0.351 hours (5.4%). Nevertheless, whether the financial investment required to achieve this performance boost can be justified, is still an open question.

With respect to resource utilization, the nurse practitioners are the bottleneck when two are present because they have the highest utilization among all providers. When a 3rd is added, the bottleneck switches to the physical therapist and the care planner whose utilizations are now over

70%. In light of these statistics, adding a 4th nurse practitioner cannot be justified.

Table 3.7: Results for different appointment templates

	VBFI-3		VBFI-4	
Metrics	Mean[1]	HW[2]	Mean	HW
Step 1 closing time	4.147	0.023	3.990	0.026
Step 2 closing time	4.149	0.023	4.005	0.025
Feasible rate	800/800		800/800	
Waiting time	0.291	0.009	0.351	0.010
Service time	1.080	0.005	1.080	0.005
Time in clinic	1.371	0.013	1.432	0.014
Waiting time (new)	0.307	0.010	0.375	0.011
Service time (new)	1.162	0.005	1.162	0.005
Time in clinic (new)	1.469	0.014	1.536	0.015
Waiting time (follow-up)	0.216	0.012	0.249	0.014
Service time (follow-up)	0.779	0.007	0.779	0.007
Time in clinic (follow-up)	0.995	0.014	1.027	0.016
Fraction above closing time	0.134	0.024	0.086	0.020

[1] All times in hours; the statistics are all Step 2 results except for the Step 1 closing time.

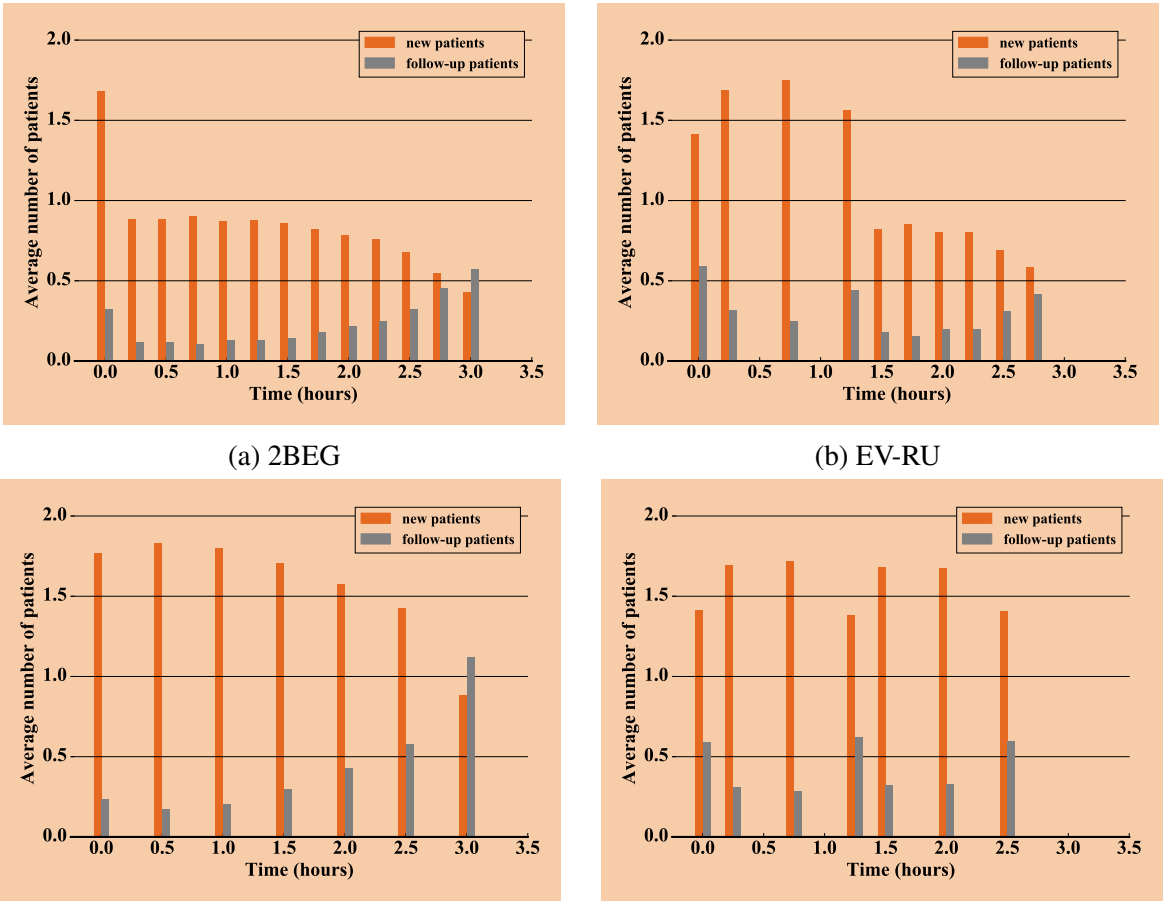
[2] Half width of a 95% confidence interval.

Table 3.8: Resource utilization

	VBFI-3		VBFI-4	
Metrics	Mean	HW	Mean	HW
Nurse practitioner 1	0.570	0.005	0.582	0.005
Nurse practitioner 2	0.508	0.005	0.528	0.005
Nurse practitioner 3	0.416	0.005	0.440	0.005
Surgeon	0.502	0.009	0.521	0.009
Physical therapist	0.709	0.010	0.734	0.010
Nutritionist	0.460	0.011	0.477	0.012
Care planner	0.710	0.005	0.736	0.005
Room	0.578	0.003	0.620	0.003

3.5.3.4 Appointment Rules

For the joint pain IPU, follow-up patients represent roughly 25% of the flow. In several recent studies, it has been shown that ordering the patient in the schedule by type can improve clinic performance (e.g., see Bosch and Dietz 2000; White et al. 2011). In this section, we propose several rules that derive from our observations of arrival patterns associated with each template for the original case with two nurse practitioners. Figure 3.8 contains four graphs that plot the average number of patients in each of the two groups who arrive at the beginning of each 15-minute interval. The graphs were constructed using the same data set that provided the computational results in Table 3.5. In this part of the analysis, our objective is to gain insight into how the model chooses appointment slots for new vs. follow-up patients under the various templates.



(a) 2BEG

(b) EV-RU

(c) VBFI-1

(d) VBFI-2

Figure 3.8: Average number of new and follow-up patients for different templates for the case with 2 nurse practitioners

Since the ratio of follow-up to new patients is 3:11, statistically, the expected number of follow-ups at each time point is the total number of patients multiplied by 3/14. By comparing the average number of follow-ups at each time point with the expected number, we can find the time slots when they have a high chance of being scheduled to arrive. For example, the expected number of follow-up patients at $t = 0$ for EV-RU is $2 \cdot 3/14 = 3/7$. In our experiments, we found that the average number of follow-ups that arrive at $t = 0$ for EV-RU is around 0.6, which is greater than 3/7. Therefore, we say the follow-up patient has a higher chance of being scheduled to arrive at

$t = 0$ for EV-RU than might be expected. Similar analysis can be done for the other time points and templates.

The following patterns appear in the graphs in Figure 3.8.

Pattern 1: A follow-up patient has a high chance of arriving at the beginning of the session.

There are explanations for this pattern: (i) Follow-up patients usually have shorter service times than new patients. Starting with one new patient and one follow-up will generally result in the latter finishing the nurse practitioner visit sooner and then moving on to his next provider. Thus, the next provider will be engaged sooner than if both patients at $t = 0$ were from the same group. Moreover, when the new patient finishes his visit with the nurse practitioner, if he is required to see the same provider as the follow-up, then his wait will likely be shorter; (ii) The difference in expected service times between the first two patients creates a staggered flow with respect to downstream providers. This tends to reduce congestion as well as the clinic closing time.

Rule 1: Schedule both a follow-up patient and a new patient at $t = 0$.

Pattern 2: When there are three or more patients scheduled at two successive time points, one of them is a follow-up patient.

At most time points, only a single patient is scheduled to arrive. At some time points in some templates, though, the patient flow can be high. In template EV-RU, for example, the total number of new and follow-up patients who arrive at successive time points $t = 1.25$ and $t = 1.5$ is 3; for VBFI-2, the total number who arrive at $t = 1.25$ and $t = 1.5$ is 4. In such cases, congestion is likely leading to long queues in front of the providers. By scheduling a follow-up patient to arrive at those time points with high inflow, the likelihood of congestion will be reduced because follow-ups typically spend less time with providers.

The second reason to schedule a follow-up patient to arrive at time points where the patient inflow is more than 2 is that all rooms are likely to be occupied. Again, follow-up patients usually

spend less time with providers, and so will spend less time in the clinic. This will help limit queuing for rooms.

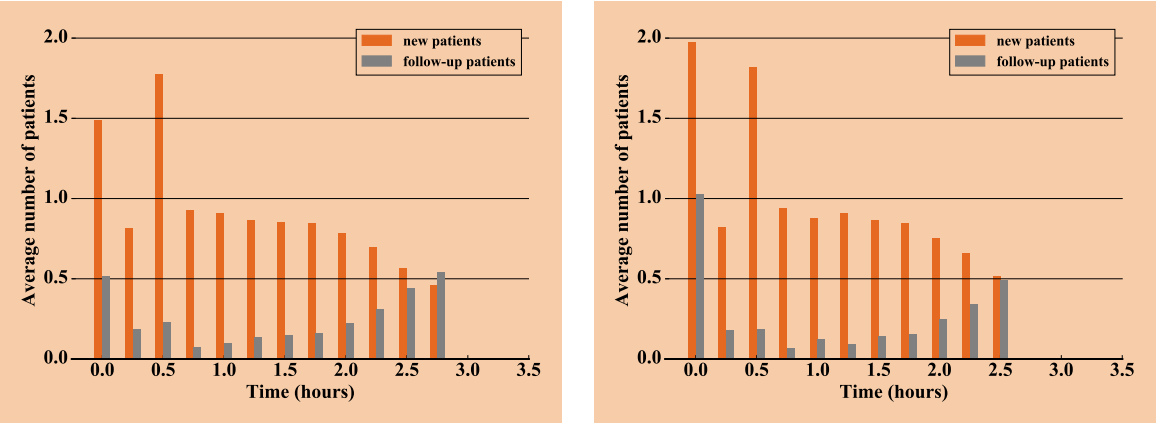
Rule 2: Embedded in the statement of Pattern 2.

Pattern 3: A follow-up patient has a high chance of arriving at the end of the session.

Pattern 3 appears in the results for all four templates. This can be explained as follows. Assume that there are 13 patients in the system and queues exist for all providers other than the two nurse practitioners. Consider the extreme case where the 14^{th} arrival is a new patient who is to be seen by all five providers. In this scenario, it is likely that the care planner has already finished her consultation with the first 13 patients before the 14^{th} patient finishes with her fourth provider. The idle time between the 13^{th} and 14^{th} patient has the effect of delaying the clinic closing time. If a follow-up patient is the last to arrive, however, it is less likely that the care planner will have finished consulting with the previous 13 patients because service times for follow-up patients are less than for new patients.

Rule 3: The last appointment should be a follow-up patient.

To check the robustness of the above patterns, an additional set of experiments was conducted to determine whether they still hold for the case with 3 nurse practitioners. The results are depicted in Figure 3.9. Indeed, Patterns 1 and 3 are still present in Figure 3.9 but Pattern 2 has disappeared. The absence of Pattern 2 is a consequence of increased capacity due to the additional nurse practitioner. Therefore, even when 3 patients are scheduled to arrive at two successive time points, there will be little if any queueing in front of any of the nurse practitioners. Hence, there is no need to schedule a follow-up patient at either time point to improve flow.



(a) VBFI-3 (b) VBFI-4

Figure 3.9: Average number of new and follow-up patients for different templates for 3 nurse practitioners

Of course, it may not be possible in practice to fully adhere to these rules due to requests for specific appointment times, provider availability, or the random nature of the patient mix. However, they do provide some level of insight and guidance for improving clinic efficiency. In our experience, outpatient scheduling is typically done on a first-come, first-served basis without taking into account patient type.

3.6 Summary and Conclusion

The complexity of patient flow in multi-provider clinics like IPUs underscores the need for a considered approach to appointment scheduling to maximize the use of available resources while assuring high levels of customer satisfaction. In this paper, we first proposed a new model for the extended open shop problem, and then specialized it to an IPU in which multiple patient and provider types have to be coordinated over the day. For the deterministic version of the problem, we developed a two-step method that provides solutions for 10 patients within 4% of optimality on average. These results were derived by analyzing a wide-range of scenarios reflecting operations of

the joint pain IPU at the Dell Medical School. A two-stage integer stochastic optimization model was then presented that more realistically represents actual patient-provider interactions. The two-step method was again used to solve the wait-and-see problem and several versions of the expected value problem. All instances contained 14 patients. The average optimality gap was less than 5% for the WS problem and less than 2% for the EV variants. Our ultimate goal has been to determine an appointment template that can be used to schedule new and follow-up patients over half-day sessions.

The results from our experimental design indicated that the templates derived from the proposed methodology provide good performance with respect to minimizing a combination of clinic closing time and patient waiting time. The relatively less aggressive templates (i) VBFI-1 (variable block/fixed interval), which allows a different number of patients to be assigned at each time point as long as they are separated by the same fixed interval, and (ii) 2BEG, where two patients are scheduled at the beginning of the session and then a single patient at fixed intervals thereafter, are preferable if patient waiting time is the clinic's primary metric. The more aggressive template EV-RU (expected value-rounded up) is more effective when the clinic closing time is of primary importance. We also observed arrival patterns by patient type for each template, and proposed several scheduling rules based on the insights gained. For example, one follow-up and one new patient should be scheduled to arrive at the beginning of the day, and one follow-up at the end. In general, similar patterns were observed in two of the three cases when we increased the number of nurse practitioners from 2 to 3. Collectively, these results have provided the foundation for designing the Dell Medical School joint pain IPU schedule.

One limitation of our model is that it does not account for the stochasticity of the arrival process. When patients depart from their scheduled appointment times by arriving early or late, the result is more uncertainty, which can lead to increased system congestion, longer queues and sojourn

times, and later closing times. The greater the uncertainty, the greater the disruption to the planned schedule. A second limitation of our work is that we did not consider patient no shows. Because of the need to coordinate multiple providers and patient types in an IPU, any disruptions in the flow can create measurable inefficiencies in clinic operations. When we began our study, we did not have the necessary data to postulate no-show probabilities for any of the six patient types because we were designing a new clinic. Rather than guessing we decided to assume that all patients arrive for their appointment on time. This allowed us to design templates for the ideal case. Further investigation and data collection are needed to determine the most effective way of dealing with no shows. Existing approaches typically resort to overbooking or shortening appointment slots to reduce the negative consequences of absent patients. However, there is no standard way of implementing either of these ideas that reliably minimizes the disruption to the system.

Chapter 4

Robust Optimization with Order Statistic Uncertainty Set

4.1 Introduction

In many optimization problems, the decision maker needs to make decisions in the presence of uncertainty. Stochastic optimization has long been used to find optimal solutions in such settings. Specifically, the random quantities are assumed to follow some probability distributions, which leads to either a random objective function, or random constraints, or both. A practical challenge of using the stochastic optimization approach is that the distributions of the random parameters are usually unknown and difficult to infer from the data. In addition, stochastic optimization models impose significant computational burden. As a result, approximation procedures are often used — see Birge and Louveaux (2011, Chapter 8 - 10).

Distributionally robust optimization is an alternative approach that hedges against distributional uncertainty. In that approach, the distributions of unknown parameters are assumed to lie in predefined distributional sets. A distributional set may be determined either by the moments of the unknown distribution (such as in Delage and Ye 2010 and Popescu 2007) or by a statistical distance measure (such as in Klabjan et al. 2013 and Gao et al. 2017). In many cases, distributionally robust optimization problems can be solved by exploiting the techniques of conic quadratic or semi-definite programming.

Another popular approach is robust optimization. Rather than model random quantities as having known distributions or distributions that belong to a set, the robust optimization model aims to find a solution that achieves the best performance of the objective function while remaining feasi-

ble for any realization (scenario) of the uncertain quantities within an *Uncertainty Set*. If properly constructed, the uncertainty set contains relatively more likely scenarios, and the decision maker may find it more economical to develop contingency plans to deal with the scenarios excluded from the uncertainty set. Robust Optimization (RO) is particularly attractive when uncertainty characterization via a probability distribution is unreliable. In this paper, we focus on robust optimization, and propose a new uncertainty set.

4.1.1 The Robust Optimization Model

Consider the standard linear optimization problem given below:

$$\max_{\mathbf{x}} \ \mathbf{c}'\mathbf{x} \tag{4.1a}$$

$$\text{s.t.} \sum_{j \in J} a_{ij} x_j \leq b_i, \ \forall i \tag{4.1b}$$

$$\underline{\mathbf{x}} \leq \mathbf{x} \leq \overline{\mathbf{x}}. \tag{4.1c}$$

where J is the index set of the vector $\mathbf{x}$ and also denotes its dimension. Bold-faced letters denote vectors and upper case letters denote random variables. Letters that are both bold face and upper case denote either matrices, or sets, or vectors of random variables. The prime notation for vectors denotes transpose. Suppose the decision maker is uncertain about the values of a_{ij}s. Consistent with our notational scheme, we denote those variables by A_{ij}s. The decision maker aims to find a solution that not only has a high objective value but also ensures feasibility of constraint (4.1b) with a specified probability. For example, to ensure the feasibility of constraint (4.1b) with probability p_i, we can use the following chance constraint.

$$\text{Prob}\left(\sum_{j \in J} A_{ij} x_j \leq b_i \right) \geq p_i, \ \forall i, \tag{4.2a}$$

The optimization problem involving chance constraints is generally hard to solve (see Yang and Xu 2016). Probabilistic feasibility of constraints can also be achieved with the RO model – see Ben-Tal and Nemirovski (2000) and Bertsimas et al. (2011a).

Uncertainty Modelling in the RO model. As in the robust optimization framework presented in Bertsimas and Sim (2004), we assume that the random variable A_{ij} follows an unknown but symmetric distribution, and A_{ij} can take any value in the range $[a_{ij} - \hat{a}_{ij}, a_{ij} + \hat{a}_{ij}]$. We transform the random variable A_{ij} into an uncertainty-level random variable $Z_{ij} \in [0,1]$ such that $Z_{ij} = |A_{ij} - a_{ij}|/\hat{a}_{ij}$, and we have $\mathbf{Z}_i \in [0,1]^J$, where J is the index set of the vector $\mathbf{x}$ and also denotes its dimension. Henceforth, whenever random variables are mentioned, we mean the random variables Z_{ij}s. Note that our approach can apply to the case where the random variables are unbounded, but we discuss the bounded case, which matches some closely related existing approaches, e.g., the budget uncertainty set. That being said, it will be straightforward to generalize our approach to the unbounded case.

The robust formulation for the linear program (4.1) with the uncertainty set $\mathscr{U}$ can be written as follows:

$$\max_{\mathbf{x}} \ \mathbf{c}'\mathbf{x} \tag{4.3a}$$

$$\text{s.t.} \ \sum_j a_{ij} x_j + \max_{\mathbf{Z}_i \in \mathscr{U}} \sum_{j \in J_i} \hat{a}_{ij} \cdot |x_j| \cdot Z_{ij} \leq b_i, \ \forall i \tag{4.3b}$$

$$\underline{\mathbf{x}} \leq \mathbf{x} \leq \overline{\mathbf{x}}, \tag{4.3c}$$

The subproblems $\beta_i(\mathbf{x}, \mathscr{U}) = \max_{\mathbf{Z}_i \in \mathscr{U}} \sum_{j \in J_i} \hat{a}_{ij} \cdot |x_j| \cdot Z_{ij}$ guarantees feasibility of the constraint i for any realization of $\mathbf{Z}_i$ that lies within the uncertainty set $\mathscr{U}$. Ben-Tal et al. (2009) in Section 1.2.1 of their paper show that one can always reformulate the above joint uncertainty set $\mathscr{U}$ to be "constraint-wise". Therefore, we will drop the constraint index i and focus on an arbitrary

constraint. As we will discuss in Section 4.3.1, the above robust formulation (4.3) is consistent with the framework presented in Bertsimas and Sim (2004). In the following, we review previous works on uncertainty set characterization, and then discuss uncertainty set design and our contribution.

4.1.2 Common Uncertainty Sets

Table 4.1 summarizes the most common uncertainty sets that have been studied in the RO literature. The table also includes the distributional information that each uncertainty set utilizes, and identifies the parameter that may be selected to adjust its size, which controls its probabilistic guarantee for constraint feasibility.

Table 4.1: Summary of uncertainty sets

Uncertainty set	Definition	Distributional information	Parameter to control the size
Interval uncertainty set	$\mathscr{U}^I = \left\{\mathbf{Z} : 0 \leq Z_j \leq 1, \forall j\right\}$	range	None
Budget uncertainty set	$\mathscr{U}^B = \left\{\mathbf{Z} : \sum_{j=1}^{J} Z_j \leq \tau, 0 \leq Z_j \leq 1, \forall j\right\}$	range	τ
Ellipsoidal uncertainty set	$\mathscr{U}^Q = \left\{\mathbf{Z} \in \mathbb{R}^J : \mathbf{Z}'\mathbf{\Sigma}^{-1}\mathbf{Z} \leq \gamma^2\right\}$	variance & covariance	γ
Demand uncertainty set	$\mathscr{U}^D = \left\{\mathbf{Z} \in \mathbb{R}^J : \left\lvert\frac{\sum_{j\in S} Z_j}{\lvert S\rvert^{1/\alpha}}\right\rvert \leq \Gamma, \forall S \subseteq J\right\}$	variance	Γ
Tail uncertainty set	$\mathscr{U}^T = \left\{\mathbf{Z} : \exists \mathbf{q} \in \mathbb{R}_+^N \text{ s.t. } \mathbf{Z} = \sum_{n=1}^{N} q_n \mathbf{z}^n, \mathbf{1}'\mathbf{q} = 1, q_n \leq \frac{1}{N(1-\alpha)}, n = 1, \cdots, N\right\}$	tail average	α

The interval uncertainty set. The interval uncertainty set (also known as the box uncertainty set) can be found in Ben-Tal and Nemirovski (2000). It offers a high protection level, but it tends to be conservative because all the random variables Z_js in the optimal solution are set to 1. That is, it finds the best solution for the worst possible realization of the unknown parameters. Bertsimas et al. (2018, Section 6) improved this approach by limiting the uncertainty in each dimension with lower and upper bounds.

The budget uncertainty set. The budget uncertainty set, introduced in Bertsimas and Sim (2004), is the first polyhedral uncertainty set that can control the level of conservativeness for the RO model (controlled by the the parameter τ). The idea is to impose the budget constraint on the sum of the all random variables Z_js, which prevents all random variables from taking the extreme value of 1.

The ellipsoidal uncertainty set. The ellipsoidal uncertainty set (Ben-Tal and Nemirovski 1998 and El Ghaoui et al. 1998) is motivated from the standard deviation formula, which results in the quadratic form. The matrix $\mathbf{\Sigma}^{-1}$ is the variance-covariance matrix for random variables Z_js.

The demand uncertainty set. The demand uncertainty set is inspired by the Generalized Central Limit Theorem (GCLT), which states that the limiting sum of independent random variables is asymptotically distributed according to a stable distribution (see Bandi and Gupta 2019). In Table 4.1, $|S|$ stands for the cardinality of the set S, which is an arbitrary subset of the set J. The parameter α is the tail coefficient and usually satisfies $1 < \alpha \leq 2$ (Bandi et al. 2015). To eliminate the extreme scenarios, the demand uncertainty set restricts the sum of the uncertain variables to be within a certain range. Some researchers have imposed such restrictions on a limited number of the possible subsets of J (Bandi et al. 2015 and Bandi and Bertsimas 2014); whereas others imposed restrictions on all possible subsets (Bertsimas et al. 2011b and Bertsimas and Bidkhori 2015). In the latter case, we have $2^J - 1$ constraints for J random variables, which grow exponentially.

The tail uncertainty set. The tail uncertainty set consists of the convex hull of all the centroids of any $N(1-\alpha)$ points out of N points in the sampled data $\mathbf{z}^1, \mathbf{z}^2, \cdots, \mathbf{z}^N$ (Bertsimas et al. 2011a). The decision variables q_ns serve as linear weights to construct the linear combination of all data points. It is a special case to the uncertainty sets proposed in Bertsimas and Brown (2009) using risk theory. The tail uncertainty set is an attractive way to characterize uncertainty if the decision maker's risk preference corresponds to the conditional value-at-risk (CVaR) measure, which may

limit its application in more general settings. Moreover, because the tail uncertainty set introduces the same number of decision variables q_ns as the number of samples in the data, it makes the RO model difficult to solve for large datasets. In some of our numerical experiments, the RO model using the tail uncertainty set is even harder to solve than using the non-linear ellipsoidal uncertainty set.

4.1.3 Principles for the Design of Uncertainty Sets

The choice of the uncertainty set is a key consideration in utilizing the RO approach. The uncertainty set in the RO model determines the trade-off between the two conflicting goals: good objective value and high probability of constraint feasibility. The balance between the objective value and probabilistic guarantee of constraint feasibility depends on two aspects of the uncertainty set. The first aspect is the *size* of the uncertainty set, which is chosen by the decision maker depending on his level of conservatism. For a chosen uncertainty set, if its size gets smaller, then the objective value improves but the probability of constraint feasibility declines; the improvement in one is always at the expense of the other. The second aspect of the uncertainty set is the *geometric flexibility*. The performance of the objective function and protection level can both be improved if the uncertainty set contains regions of more likely uncertain scenarios and excludes the extremely unlikely ones. To achieve this, we need to design the uncertainty set with greater geometric flexibility so that we can adjust its shape to contain regions of high probability.

After we design the uncertainty set that possesses geometric flexibility, we then need to identify characterizations of uncertainties that can guide us to adjust the shape of the uncertainty set. Such characterization may stem from two sources. One source can be the data-free and distribution-free properties of random variables. Typically, the properties emanate from general statistical knowledge of random variables, which requires no input from any sample data, and as few assumptions as

possible about the data generating process. The second source is the specific information relevant to the particular problem setting, which may be derived from either historical data or institutional knowledge. Existing uncertainty sets have utilized different kinds of information, which often appear as parameters in the formulations; for example, the range in the budget uncertainty set, or the mean and variance in the ellipsoidal uncertainty set. However, these statistics contain very limited information and may lose some useful distributional information, e.g., the peakedness of the distribution. Therefore, it is important to conceive uncertainty sets that can incorporate richer information from data.

4.1.4 Our Contribution

The focus of this study is to explore the characterization of random variables and utilize it to design a new uncertainty set. We seek to answer the following questions: What are the data-free and distribution-free statistical characteristics of the collective behavior of random variables that may be utilized to refine the uncertainty set? How can we design an uncertainty set that captures rich distributional information (richer than the mean and the variances/covariances), while still resulting in a linear programming formulation? Is it possible to construct an uncertainty set that offers the ability to adjust the level of uncertainty in each dimension separately rather than a single parameter that affects all dimensions in the same way? Our main results are as follows:

1. We use the Probability Integral Transform to show that if the random variables Z_js are continuous and mutually independent of each other, then the order statistics of the cumulative distribution functions (CDFs) of Z_js follow the Beta distribution. As a result, each order statistic of the CDFs of random variables Z_js has a confidence interval within the range $[0,1]$ for a given probability. Based on this data-free distribution-free property of CDFs of random variables Z_js, we construct a new order statistic uncertainty set by imposing constraints on

order statistics of the CDFs of random variables Z_js.

2. To embed the CDFs of random variables in the formulation of the order statistic uncertainty set, we utilize the quantiles of random variables, which carry rich distributional information of random variables. Because the order statistics of the CDFs of Z_js have $J!$ possible outcomes, the constraints for them imply $J!$ implicit linear constraints. In order to develop a tractable linear formulation for the $J!$ implicit linear constraints, we adopt the formulation of the assignment problem.

3. We demonstrate the geometric flexibility of the order statistic uncertainty set by showing that it reduces to either the interval uncertainty set, or the budget uncertainty set, or the demand uncertainty set if its parameters are selected appropriately. This shows that the order statistic uncertainty set has a greater modeling power because it incorporates these three uncertainty sets as special cases. The new uncertainty set also captures richer information than other existing uncertainty sets because it utilizes the quantiles of distributions to characterize uncertainties.

4. We provide a probabilistic guarantee for the constraint feasibility of the solution from the RO model with the new uncertainty set. We also present statistical methods to estimate the parameters used in our uncertainty set for those instances in which data are available. Finally, we apply our uncertainty set and several competing characterizations of the uncertainty set to both synthetic data and real data sets to compare and contrast their relative performance.

The outline of the paper is as follows. In Section 4.2, we present the motivation to construct the order statistic uncertainty set and a linear formulation of the RO model with the order statistic uncertainty set. In Section 4.3, we analyze the advantages of the order statistic uncertainty set, and show that three existing uncertainty sets may be viewed as special cases of the order statistic

uncertainty set. In Section 4.4, we derive the probabilistic bound for constraint feasibility for the RO model with the order statistic uncertainty set, and discuss how one may estimate its parameters. In Section 4.5, we apply the RO models to solve portfolio optimization with shortfall constraints and compare the performance of the order statistic uncertainty set and other existing uncertainty sets. We conclude in Section 4.6.

4.2 The Order Statistic Uncertainty Set

In this section, we first study a property of random variables Z_js using the Probability Integral Transformation, based on which we construct the order statistic uncertainty set. In Section 4.2.2, we present a linear formulation of the RO model with the new uncertainty set.

4.2.1 Preliminaries

Suppose the random variables Z_js are continuous and independently distributed in the range $[0,1]$, each following an arbitrary continuous distribution with unknown cumulative distribution function F_j. Let $U_j = F_j(Z_j), \forall j \in J$. It is well-known that U_j's are uniformly distributed over $[0,1]$ (see Roussas 1997, Section 9.4). Denote the order statistics of U_js as $U_{(1)},\ldots,U_{(J)}$, which is the rearranged sequence of U_js with k-th order statistic $U_{(k)}$ being the k-th smallest among them. Although the random variable U_j follows Unif$(0,1)$ distribution, $U_{(k)}$ does not follow the uniform distribution. The probability density function of $U_{(k)}$ follows $Beta(k, J+1-k)$ distribution (see Gut 2009, Chapter 4.1). The mapping from $\{Z_1,\cdots,Z_J\}$ to $\{U_{(1)},\ldots,U_{(J)}\}$ is illustrated in Figure 4.1.

Before describing our approach, we first explain the key idea with the help of an example with $J = 20$. Figure 4.2 shows the $Beta(k, J+1-k)$ distribution of $U_{(k)}$s, $\forall k = 1,\cdots,20$ from left to right. There are two observations worth noting.

1. If k is small, the distribution of $U_{(k)}$ tends to be right skewed, which means the $U_{(k)}$ variable

Variables	Distribution
$Z_1, Z_2, \dots, Z_J$	Z_js are independent with arbitrary continuous distributions
↓ via cdf transformation	
$U_1 = F_1(Z_1), U_2 = F_2(Z_2), \dots, U_J = F_J(Z_J)$	$U_j \sim \text{Uniform}(0,1), \forall j$
↓ via ordering from smallest to largest	
$U_{(1)}, U_{(2)}, \dots, U_{(J)}$	$U_{(k)} \sim Beta(k, J+1-k)$

Figure 4.1: Transformations of variables.

tends to be small. As k increases, the distribution of $U_{(k)}$ gets more skewed to the left. Most $U_{(k)}$s are extremely unlikely to be either 0 or 1.

2. Each order statistic $U_{(k)}$ has an interval strictly smaller than $[0,1]$, over which the area under its pdf is close to 1. For example, the area under the 8th order statistic's pdf (the solid line) over the interval $[0.05, 0.85]$ is 0.999997, which is almost 1! This illustrates that the uncertainty characterization with either the box or the budget uncertainty sets is too extreme because they have at least $J-1$ random variables Z_js that are equal to either 0 or 1.

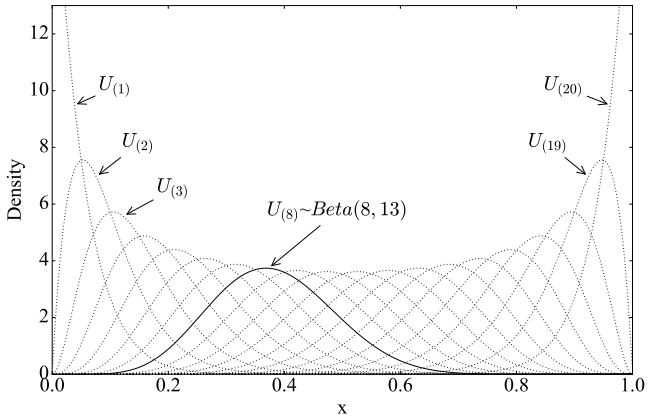


Figure 4.2: Probability density functions of order statistics $U_{(k)}$s for $J = 20$.

From above, we see that regardless of the unknown distributions F_js of the random variables Z_js, the order statistics of $F_j(Z_j)$s always follow the Beta distribution, which allows us to explore the following intrinsic property of the random variables. Given any $\varepsilon'_k \geq 0$, we should be able to find the lower limit $u^l_{(k)} > 0$ for $U_{(k)}$, such that $[u^l_{(k)}, 1]$ is a confidence interval with $1-\varepsilon'_k$ confidence. Similarly, we can find the upper confidence limit $u^u_{(k)} < 1$ with $1-\varepsilon_k$ confidence. Define the quantile function $Q^t_k = \inf\{x : I_x(k, J+1-k) \geq t\}$, where $I_x(k, J+1-k)$ is the CDF for $Beta(k, J+1-k)$ distribution. It then holds that $Prob\big(U_{(k)} \leq Q^t_k\big) \geq t$. Denote $\boldsymbol{\epsilon'}$ as the vector of values $\varepsilon'_1, \varepsilon'_2, \cdots, \varepsilon'_J$, and $\boldsymbol{\epsilon}$ of values $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_J$, where $0 \leq \varepsilon'_j, \varepsilon_j \leq 1, \forall j \in J$. Next, we construct the uncertainty set $\mathscr{U}'(\boldsymbol{\epsilon'}, \boldsymbol{\epsilon})$ in terms of the order statistics of the CDFs of random variables as follows:

$$\mathscr{U}'(\boldsymbol{\epsilon'}, \boldsymbol{\epsilon}) = \Big\{ \mathbf{Z} : F_j(Z_j) = U_j, \text{ and } Q^{\varepsilon'_k}_k \leq U_{(k)} \leq Q^{(1-\varepsilon_k)}_k, \ \forall k, j \in J \Big\}.$$

The above uncertainty set contains high-density regions for the order statistics of CDFs F_js of random variables Z_js. In other words, the ranges of Z_js may be restricted such that the order statistics $U_{(k)}$s belong to the uncertainty set $\mathscr{U}'(\boldsymbol{\epsilon'}, \boldsymbol{\epsilon})$. Note that $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(J)}$ is always implied by definition. Although ε'_k can be any value in the range $[0,1]$, we only need to consider ε'_ks such that $Q^{\varepsilon'_{k-1}}_{k-1} \leq Q^{\varepsilon'_k}_k$, for $k = 2, 3, \cdots, J$. To argue for this, we first assume that there exists a k_0 such that $Q^{\varepsilon'_{k_0-1}}_{k_0-1} > Q^{\varepsilon'_{k_0}}_{k_0}$. If $U_{(k_0)} = Q^{\varepsilon'_{k_0}}_{k_0}$ holds, then any $U_{(k_0-1)}$ that satisfies $U_{(k_0-1)} \geq Q^{\varepsilon'_{k_0-1}}_{k_0-1}$ would violate the constraint $U_{(k_0-1)} \leq U_{(k_0)}$. For a similar reason, we also assume $Q^{(1-\varepsilon_{k-1})}_{k-1} \leq Q^{(1-\varepsilon_k)}_k$, for $k = 2, 3, \cdots, J$.

Remark 1. *The above uncertainty set $\mathscr{U}'(\boldsymbol{\epsilon'}, \boldsymbol{\epsilon})$ is fundamentally different from the uncertainty set in Bertsimas et al. (2018, Section 6). We illustrate this with the special case where $F_j \equiv F, \forall j \in J$. For simplicity, we also assume the function F is strictly increasing, so that F^{-1} exists. Denote the*

kth order statistics of Z_js as $Z_{(k)}$. Then $\mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})$ becomes

$$\left\{\mathbf{Z} : Q_k^{\varepsilon'_k} \leq F(Z_{(k)}) \leq Q_k^{(1-\varepsilon_k)},\ \forall k \in J\right\},$$

which can be rewritten as:

$$\left\{\mathbf{Z} : F^{-1}\left(Q_k^{\varepsilon'_k}\right) \leq Z_{(k)} \leq F^{-1}\left(Q_k^{(1-\varepsilon_k)}\right),\ \forall k \in J\right\}.$$

The uncertainty set in Bertsimas et al. (2018, Section 6) can be described as follows:

$$\left\{\mathbf{Z} : lb_j \leq Z_j \leq ub_j,\ \forall j \in J\right\},$$

where lb_j and ub_j are lower and upper limits on Z_j, and the shape of the uncertainty set is a hyperrectangle. In contrast, our uncertainty set is not directly defined on Z_js but on their order statistics. As a result, it is no longer a hyperrectangle and in fact is generally non-convex as we will show in Section 4.2.2. □

Earlier, we defined $\beta\left(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\right) = \max_{\mathbf{Z} \in \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})} \sum_{j \in J} \hat{a}_j \cdot |x_j| \cdot Z_j$. The following characterization of $\mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})$ helps to reformulate the RO model with the uncertainty set $\mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})$.

Proposition 12. *Given $\boldsymbol{\epsilon}'$, $\boldsymbol{\epsilon}$, and a fixed* **x**, *$U_{(k)} = Q_k^{(1-\varepsilon_k)}, \forall k$ maximizes $\beta\left(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\right)$.*

Proposition 12 states that the order statistics of $F_j(Z_j)$s are at their upper bounds in the optimal solution to $\beta\left(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\right)$. Note that Z_j's coefficient $\hat{a}_j \cdot |x_j|$ is always non-negative, then in order to maximize $\beta\left(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\right)$, Z_j should be as large as possible. Because F_j is non-decreasing, $F_j(Z_j)$ should also be as large as possible. As a result, each order statistic $U_{(k)}$ of $F_j(Z_j)$s should be as large as its upper bound $Q_k^{(1-\varepsilon_k)}$. Based on this property, we define the order statistic uncertainty

set in the following equivalent way:

$$\mathscr{U}^{OS}(\boldsymbol{\epsilon}) = \Big\{\mathbf{Z} : F_j(Z_j) = U_j, \text{ and } U_{(k)} \leq Q_k^{(1-\varepsilon_k)}, \ \forall k, j \in J\Big\}.$$

Because $U_{(k)} = Q_k^{(1-\varepsilon_k)}$ should also hold in the optimal solution to $\beta\big(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\big)$, we have $\beta\big(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\big) = \beta\big(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\big)$. We emphasize that all the properties we presented above only rely on the assumption that the continuous distributions of Z_js are independent; these properties are distribution-free because they hold regardless of the distributions F_js of the random variables of Z_js; they are also data-free because they are not based on any information extracted from data.

4.2.2 Robust Optimization with Order Statistic Uncertainty Set

The order statistic uncertainty set $\mathscr{U}^{OS}(\boldsymbol{\epsilon})$ is intractable in its current form for three reasons. The first reason is that it is not directly defined on variable Z_js, but in the space of CDFs of Z_js. Another reason is that there are $J!$ permutations of $F_j(Z_j)$s for all possible outcomes of $U_{(k)}$s, which makes reformulating it even challenging. The third reason has to do with its nonconvexity as stated in the following proposition.

Proposition 13. *If there exist k_1 and k_2 ($1 \leq k_1 < k_2 \leq J$) such that $Q_{k_1}^{(1-\varepsilon_{k_1})} \neq Q_{k_2}^{(1-\varepsilon_{k_2})}$, then the uncertainty set $\mathscr{U}^{OS}(\boldsymbol{\epsilon})$ is not convex.*

In what follows, we apply the formulation for the assignment problem to develop a linear formulation for the RO model with the order statistic uncertainty set $\mathscr{U}^{OS}(\boldsymbol{\epsilon})$. Let q_{jk} be the quantile value: $q_{jk} = \sup\{x : F_j(x) \leq Q_k^{(1-\varepsilon_k)}\}, \forall j,k \in J$. The value q_{jk} can be viewed as the Z_j's quantile of order $Q_k^{(1-\varepsilon_k)}$, and we discuss how to estimate it in Section 4.4.2. The following proposition provides a tractable formulation for the problem $\beta\big(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\big)$.

Proposition 14. *For a fixed* $\mathbf{x}$*, the* $\beta\left(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\right)$ *function may be obtained by solving the following linear optimization problem:*

$$\max_{\eta} \ \sum_{j\in J} \hat{a}_j |x_j| \cdot \left(\sum_{k\in J} q_{jk}\eta_{jk} \right) \tag{4.4a}$$

$$s.t. \ \sum_{k} \eta_{jk} = 1, \forall j \in J \tag{4.4b}$$

$$\sum_{j} \eta_{jk} = 1, \forall k \in J \tag{4.4c}$$

$$0 \leq \eta_{jk} \leq 1, \forall j,k \in J. \tag{4.4d}$$

The problem (4.4) in Proposition 14 is the linear relaxation of the maximum weight assignment problem, which is known to have an integer optimal solution. If $\eta_{jk} = 1$, then $\hat{a}_j|x_j|$ is assigned to q_{jk}, which implies $Z_j = q_{jk}$ and $F_j(Z_j) = Q_k^{(1-\varepsilon_k)}$. Because the integer optimal solution of the assignment problem is a bijective mapping, the set $\left\{F_j(Z_j), \forall j \in J\right\}$ will be mapped to the set $\left\{Q_k^{(1-\varepsilon_k)}, \forall k \in J\right\}$. This is essentially the same as the optimal characterization $\beta\left(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\right)$ in Proposition 12.

Because the problem $\beta\left(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\right)$ is a linear program for every fixed $\mathbf{x}$, its optimal solutions are at the extreme points and we have $\beta\left(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\right) = \beta\left(\mathbf{x}, \mathbf{conv}\left(\mathscr{U}^{OS}(\boldsymbol{\epsilon})\right)\right)$, where $\mathbf{conv}\left(\mathscr{U}^{OS}(\boldsymbol{\epsilon})\right)$ is the convex hull of $\mathscr{U}^{OS}(\boldsymbol{\epsilon})$. From the proof of Proposition 14, we can know that the convex hull of $\mathscr{U}^{OS}(\boldsymbol{\epsilon})$ and feasible region of the problem (4.4) are the same. In other words, the protection region of the uncertainty set $\mathscr{U}^{OS}(\boldsymbol{\epsilon})$ expands to its convex hull.

We now study the RO Model (4.3) with uncertainty set $\mathscr{U}^{OS}(\boldsymbol{\epsilon})$. We follow the procedure in Bertsimas and Sim (2004) to reformulate the following model to a linear optimization model (we

add back the index i for the constraints in the remainder of this section).

$$\max_{\mathbf{x}} \ \mathbf{c}'\mathbf{x} \tag{4.5a}$$

$$\text{s.t.} \ \sum_j a_{ij}x_j + \max_{\mathbf{Z}_i \in \mathscr{U}_i^{OS}(\epsilon)} \sum_{j \in J_i} \hat{a}_{ij} \cdot |x_j| \cdot Z_{ij} \leq b_i, \ \forall i \tag{4.5b}$$

$$\underline{\mathbf{x}} \leq \mathbf{x} \leq \bar{\mathbf{x}}. \tag{4.5c}$$

Proposition 15. *Model* (4.5) *is equivalent to the following linear programming problem:*

$$\max \ \sum_j c_j x_j \tag{4.6a}$$

$$s.t. \ \sum_j a_{ij}x_j + \sum_{j \in J_i} (\theta_{ij} + \phi_{ij}) + \sum_{j \in J_i} \sum_{k \in J_i} \zeta_{ijk} \leq b_i, \ \forall i \tag{4.6b}$$

$$-y_j \leq x_j \leq y_j, \forall j \tag{4.6c}$$

$$\underline{x}_j \leq x_j \leq \bar{x}_j, \forall j \tag{4.6d}$$

$$\theta_{ij} + \phi_{ik} + \zeta_{ijk} \geq \hat{a}_{ij} q_{ijk} y_j, \forall j, k \in J_i, \forall i \tag{4.6e}$$

$$y_j \geq 0, \forall j \tag{4.6f}$$

$$\zeta_{ijk} \geq 0, \forall j, k \in J_i, \forall i \tag{4.6g}$$

We leverage the strong duality to obtain the linear formulation (4.6) by replacing the maximizing problem in constraints (4.5b) with the dual of problem (4.4). Because Model (4.6) requires $\mathscr{O}(J^2)$ variables and $\mathscr{O}(J^2)$ constraints, its computational complexity is slightly higher than the RO model with the budget uncertainty set which requires $\mathscr{O}(J)$ variables and $\mathscr{O}(J)$ constraints.

4.3 Comparison with Other Uncertainty Sets

In this section, we demonstrate the geometric flexibility of the order statistic uncertainty set. We first compare the order statistic uncertainty set with three uncertainty sets that have been proposed in the literature, and show that they may be viewed as special cases of the order statistic uncertainty set.

4.3.1 Comparison with the Box and the Budget Uncertainty Set

Although motivated by different statistical properties, the order statistic uncertainty set has a close relationship with the box and the budget uncertainty sets. We illustrate it with a numerical example with $J = 7$. The general structures of the Z_j values in the optimal solutions for different uncertainty sets are shown in Figure 4.3. In each figure, the values of Z_js are ordered from the smallest to the largest. In the optimal solution of the RO model with the order statistic uncertainty set, the values of Z_js are fractions $q_{j_1,1}, q_{j_2,2}, \cdots, q_{j_J,J}$, where $j_1, j_2, \cdots, j_J$ is a sequence of $1, 2, \cdots, J$. These J fractional values can be completely different from each other as shown in Figure 4.3. For any particular k, the fractional value $q_{j_k,k}$ has up to J possible outcomes because it depends on the index j_k. The values of $q_{j_1,1}, q_{j_2,2}, \cdots, q_{j_J,J}$ depend on the sequence $j_1, j_2, \cdots, j_J$, which has up to $J!$ outcomes.

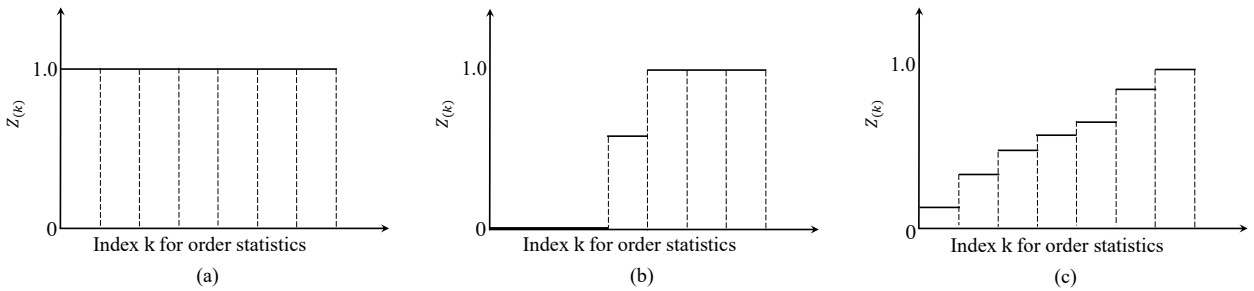


Figure 4.3: Order statistics of Z_js for different uncertainty sets — (a) box uncertainty set, (b) budget uncertainty set, (c) order statistic uncertainty set.

The box and the budget uncertainty sets are both special cases of the order statistic uncertainty set. We can obtain these two uncertainty sets by choosing specific values of the parameters of the order statistic uncertainty set.

1. The RO model with the order statistic uncertainty set reduces to the RO model with the interval uncertainty set if we choose $q_{jk} = 1, \forall k, j \in J$. The shape of the interval uncertainty set is always a hyperrectangle, which is not adjustable.

2. The RO model with the order statistic uncertainty set reduces to the RO model with the budget uncertainty set with budget τ if we choose q_{jk} as follows: $q_{jk} = 0$, if $1 \leq k \leq J - \lfloor \tau \rfloor - 1, \forall j \in J$; $q_{jk} = \tau - \lfloor \tau \rfloor$, if $k = J - \lfloor \tau \rfloor, \forall j \in J$; $q_{jk} = 1$, if $J - \lfloor \tau \rfloor + 1 \leq k \leq J, \forall j \in J$. In Appendix C.5, we prove that the RO model with the order statistic uncertainty set with such a choice of q_{jk} values is equivalent to the RO model with the budget uncertainty set with budget τ (the Problem (4) in Bertsimas and Sim 2004). In the optimal solution of the RO model with the budget uncertainty set, only one of Z_js can be a fraction, and all other Z_js are either 0 or 1 (Bertsimas and Sim 2004, Section 3). The geometric flexibility of the budget uncertainty set is limited because a small change in the value of the budget will only change the value of the fractional Z_j, but not any other Z_js.

4.3.2 Comparison with the Demand Uncertainty Set

Next we show that the demand uncertainty set that has $2^J - 1$ constraints for the Z_js can be obtained from the formulation of the order statistic uncertainty set. Recall that the demand uncertainty set is as follows:

$$\mathscr{U}^D = \left\{ \mathbf{Z} \in \mathbb{R}^J : \left| \frac{\sum_{j \in S} Z_j}{|S|^{1/\alpha}} \right| \leq \Gamma, \forall S \subseteq J \right\}.$$

In the literature α is assumed to be in the range $(1,2]$, and $\Gamma \geq 0$. The following proposition gives the optimal solution to $\beta\left(\mathbf{x}, \mathscr{U}^D\right)$ and connects the demand uncertainty set with the order statistic uncertainty set.

Proposition 16. *For a fixed* $\mathbf{x}$, $Z^*_{(k)} = \Gamma(J+1-k)^{1/\alpha} - \Gamma(J-k)^{1/\alpha}, \forall k \in J$ *maximizes* $\beta\left(\mathbf{x}, \mathscr{U}^D\right)$.

We provide the key idea of the above result. Note the objective of $\beta\left(\mathbf{x}, \mathscr{U}^D\right)$ is the sum of the pairwise products of two sequences $\hat{a}_j|x_j|$s and Z_js. Because of the rearrangement inequality (see Cvetkovski 2012, Theorem 6.1), in the maximizer of $\beta\left(\mathbf{x}, \mathscr{U}^D\right)$, the two sequences should be in the same order, i.e., both non-decreasing or non-increasing. Thus the largest $\hat{a}_j|x_j|, \forall j \in J$ should be paired with the largest $Z_j, \forall j \in J$, and the second largest of the two sequences should also be paired with each other, and so on. Because all the $\hat{a}_j|x_j|$s are non-negative, Z_j should be as large as possible in order to maximize $\beta\left(\mathbf{x}, \mathscr{U}^D\right)$. Moreover, we should make the largest of $Z_j, \forall j \in J$ to be as large as possible because it is paired with the largest $\hat{a}_j|x_j|, \forall j \in J$. The largest of $Z_j, \forall j \in J$ is restricted by the constraints with $|S| = 1$, so it should be equal to Γ. The second largest of $Z_j, \forall j \in J$ is restricted by the constraints with $|S| = 2$, so it should be equal to $\Gamma \cdot (2^{1/\alpha} - 1)$. The remaining Z_js can be analyzed similarly.

Different from the order statistic uncertainty set, the value of $Z^*_{(k)} = \Gamma(J+1-k)^{1/\alpha} - \Gamma(J-k)^{1/\alpha}$ in the demand uncertainty set does not depend on the index j of Z_js. This shows that the demand uncertainty set has less geometric flexibility than the order statistic uncertainty set because it does not capture the potential heterogeneity of the distributions of Z_js.

The following corollary provides an equivalent formulation for the demand uncertainty set. It requires J^2 continuous variables, and $J^2 + 2J$ constraints, which is much less than the $2^J - 1$ constraints in $\mathscr{U}^D$ if J is large.

Corollary 2. *For a fixed* $\mathbf{x}$*, the optimal solution to problem* $\beta\left(\mathbf{x}, \mathscr{U}^D\right)$ *is equivalent to the following*

linear optimization problem:

$$\max_{\eta} \sum_{j\in J} \hat{a}_j|x_j| \cdot \left(\sum_{k\in J} \Gamma \cdot \left(k^{1/\alpha} - (k-1)^{1/\alpha}\right) \cdot \eta_{jk} \right) \tag{4.7a}$$

$$s.t. \sum_{k} \eta_{jk} = 1, \forall j \in J \tag{4.7b}$$

$$\sum_{j} \eta_{jk} = 1, \forall k \in J \tag{4.7c}$$

$$0 \leq \eta_{jk} \leq 1, \forall j,k \in J \tag{4.7d}$$

For the problem $\beta\left(\mathbf{x}, \mathscr{U}^D\right)$, we need to find the maximum sum of the pairwise products of the sequence $\hat{a}_j|x_j|, \forall j \in J$ and the sequence $\Gamma \cdot \left(k^{1/\alpha} - (k-1)^{1/\alpha}\right), \forall k \in J$. Corollary 2 ensures that the elements of the two sequences are one-to-one paired with the assignment formulation. In the maximizer of the problem (4.7), the two sequences must be in the same order, and thus the sum of the pairwise products of the two sequences is maximized.

4.3.3 Advantages of Quantiles

Using the quantiles of the distribution to construct the order statistic uncertainty set has the following three advantages. (1) The quantile is a robust statistic and less sensitive to the extreme observations that some other statistics, e.g., the mean and the variance. (2) The uncertainty of each random variable is depicted by J quantiles, which contain richer information about the distribution of the uncertainty than the range, or the mean, or the variance. (3) Consider the extreme scenario of the uncertainty set where one of Z_js reaches its maximum Z_j^{max} and all other Z_js are equal to 0. For such extreme cases in the order statistic uncertainty set, $F_j(Z_j^{max}) = Q_J^{1-\varepsilon_J}, \forall j \in J$ hold according to the definition of the order statistic uncertainty set. Therefore, the extreme scenarios in the order statistic uncertainty set are fair for different Z_js in the sense that $F_j(Z_j^{max})$ is the same for different

js even when the distributions for Z_js are different. However, other uncertainty sets do not have this property unless for special cases, e.g., when the distributions for different Z_js are the same.

4.4 Further Analysis of the Order Statistic Uncertainty Set

In this section, we first derive the probabilistic guarantee of the order statistic uncertainty set to ensure the constraint feasibility under uncertainty. Then in Section 4.4.2, we illustrate how one may estimate the quantile values $(q_{jk}s)$ that we introduced in Section 4.2.2 from available data.

4.4.1 Probability of Constraint Feasibility

Denote the optimal solution to Model (4.5) as $\mathbf{x}^*, \beta^*$; constraint index i is dropped in this section for the ease of exposition, i.e., we study a single constraint in (4.3b). Because each order statistic $U_{(k)}$ has no more than ε_k probability of violation, we can easily prove that the order statistic uncertainty set provides at least $(1 - \sum_{k\in J} \varepsilon_k)$ probabilistic guarantee of feasibility. However, it is a rather low probabilistic guarantee. The reason is that it is not mutually exclusive for different $U_{(k)}$s to violate the uncertainty set, so the probability of violation is less than $\sum_{k\in J} \varepsilon_k$ according to the addition rule of probability, and $1 - \sum_{k\in J} \varepsilon_k$ is the lower bound for the probabilistic guarantee. In the following we prove a higher probabilistic guarantee for a special case where the random variables are independently and symmetrically distributed. The probabilistic guarantee is expressed by a formula derived in Steck (1971), which gives the probability of order statistics of the uniform distribution lying in a multi-dimensional rectangle.

Proposition 17 (This is a restatement of the first theorem in Steck 1971)**.** *Let Δ be the $n \times n$ matrix*

whose (i, j)*-th element is given as:*

$$\Delta_{ij} = \begin{cases} \left(Q_i^{(1-\varepsilon_i)}\right)^{j-i+1}/(j-i+1)!, & j-i+1 \geq 0 \\ 0, & j-i+1 < 0, \end{cases}$$

we have

$$prob\left(U_{(k)} \leq Q_k^{(1-\varepsilon_k)}, k = 1, \cdots, J\right) = J! \det[\Delta]. \tag{4.8a}$$

Proposition 18. *If* $A_j, \forall j \in J$ *are independently and symmetrically distributed in* $[a_j - \hat{a}_j, a_j + \hat{a}_j]$*, then the order statistic uncertainty set* $\mathscr{U}^{OS}(\boldsymbol{\epsilon})$ *implies a probabilistic guarantee of at least* $\frac{1}{2} + \frac{1}{2} \cdot J! \det[\Delta]$ *for the feasibility of the constraint* (4.3b).

If the conditions in the above proposition are satisfied, then the solution from the RO model with the order statistic uncertainty set $\mathscr{U}^{OS}(\boldsymbol{\epsilon})$ can ensure that the probability of the constraint (4.3b) being feasible is at least $\frac{1}{2} + \frac{1}{2} \cdot J! \det[\Delta]$. We can use the above result to determine the parameters for the order statistic uncertainty set when given a requirement of the probabilistic guarantee. If a historical dataset is available, after we solve the RO model and obtain its solution, we can evaluate an empirical posterior probabilistic guarantee. Depending on the relative magnitude of the empirical posterior probabilistic guarantee and the required probabilistic guarantee, we can adjust the size of the order statistic uncertainty set accordingly. If the empirical posterior probabilistic guarantee is smaller (larger) than the required probabilistic guarantee, then we can increase (reduce, respectively) the size of the order statistic uncertainty set.

4.4.2 Estimating Quantiles in the Order Statistic Uncertainty Set

The RO model (4.6) for the order statistic uncertainty set requires q_{jk}s as inputs. Otherwise, if we cannot obtain the quantiles of random variables, then this approach may not be applicable. In practice, if there is no historical data, decision makers may choose these parameters based on institutional knowledge. If there is data, then it is useful to know how to estimate q_{jk}s for the order statistic uncertainty set.

The parameter q_{jk} is the random variable Z_j's quantile of order $Q_k^{(1-\varepsilon_k)}$. Suppose we have N samples of Z_j denoted as $z_j^1, \cdots, z_j^N$. The simple random sampling gives the following estimation of $q_{jk} = \sup\{x : F_j(x) \leq Q_k^{(1-\varepsilon_k)}\}, \forall j,k \in J$.

$$q_{jk}^{SRS} = \max\left\{ z_j^m : \frac{\sum_{n\in N} \mathbb{1}_{z_j^n \leq z_j^m}}{N} \leq Q_k^{(1-\varepsilon_k)}, \forall m \in N \right\}. \tag{4.9a}$$

Note that the above method provides quantile estimations with discontinuities. To resolve this issue, we can apply the interpolation or smoothing techniques (see Dielman et al. 1994). The simple random sampling estimator is asymptotically normal, and the asymptotic variance could be reduced by various variance reduction approaches (see Glasserman et al. 2000), including stratified sampling, importance sampling, etc.

4.5 Numerical Experiments

In this section, we test our approach on portfolio optimization problem with shortfall constraint presented in Bertsimas et al. (2011a) using both synthetic data that we generated and real data from Kenneth French's website (see French 2019). The purpose of the test is to evaluate the relative performance of different RO approaches. Therefore, as in other RO literature, we do not account for certain aspects of standard portfolio optimization models, e.g., we do not take transaction fees

into account.

Suppose a decision maker wants to allocate one unit of asset among J portfolios. Portfolio j's return is assumed to be a random variable in range $[r_j - \hat{r}_j, r_j + \hat{r}_j]$, and can be denoted as $r_j + \rho_j \hat{r}_j,\ -1 \leq \rho_j \leq 1$. As defined earlier, $Z_j = |\rho_j|, \forall j \in J$. Suppose we invest x_j in portfolio j, and the goal is to maximize the expected return subject to the following constraint on shortfall probability.

$$\text{prob}\left(\sum_{j \in J} x_j (r_j + \rho_j \hat{r}_j) \leq s \right) \leq p_s,$$

where s is the threshold return below which the shortfall is defined to occur, and p_s is the maximum acceptable shortfall probability, i.e., the probability of return being less than s should not be greater than p_s. Instead of explicitly constraining the shortfall risk, the RO model restricts the return to be no less than the benchmark return s, for all possible realizations of ρ_js within a specified uncertainty set. Suppose we have N samples $\rho_{j1}, \cdots, \rho_{jN}$ for each ρ_j. The following RO model with uncertainty set $\mathscr{U}$ maximizes the expected return subject to the feasibility constraint.

$$\max_{\mathbf{x}} \ \sum_{n \in N} \sum_{j \in J} x_j (r_j + \rho_{jn} \hat{r}_j)/N \tag{4.10a}$$

$$\text{s.t.} \ \sum_{j \in J} x_j (r_j - Z_j \hat{r}_j) \geq s, \forall\, \mathbf{Z} \in \mathscr{U} \tag{4.10b}$$

$$\sum_{j \in J} x_j = 1, \tag{4.10c}$$

$$0 \leq x_j \leq 1, \forall j \in J \tag{4.10d}$$

For each uncertainty set $\mathscr{U}$, we can solve problem (4.10) with different values of the parameters in $\mathscr{U}$, and different solutions have different levels of expected return and different shortfall probabilities, which consist of the efficient frontier. For a given expected return r_0, we can solve a

stochastic optimization (SP) problem to obtain the solution with lowest shortfall probability. The stochastic optimization model can be formulated to a mixed integer program (MIP) using sample average approximation as follows:

$$\min_{\mathbf{x}} \sum_{n=1}^{N} \left(\mathbb{1}_{\sum_j x_j (r_j + \rho_{jn} \cdot \hat{r}_j) \leq s} \right) / N \tag{4.11a}$$

$$\text{s.t.} \sum_{n=1}^{N} \sum_{j} x_j (r_j + \rho_{jn} \cdot \hat{r}_j) / N \geq r_0, \tag{4.11b}$$

$$\sum_{j \in J} x_j = 1, \tag{4.11c}$$

$$0 \leq x_j \leq 1, \forall j \in J, \tag{4.11d}$$

In the next two sections, we present the results of RO models with different uncertainty sets and the results of the MIP formulation. All computations were performed in C++ using IBM's Concert Technology library and solved with CPLEX 12.7 on a Linux workstation with 4 Intel(R) Core(TM) i7-4790 CPU, 8 3.60GHz processors and 16 GB memory running Ubuntu 16.04. The values of ε_ks in the order statistic uncertainty set are selected to be the same for all $k \in J$, which can range from 0 to 1.

4.5.1 Experiments with Synthetic Data

In the RO model, the uncertain stock return variable in the range $[r_j - \hat{r}_j, r_j + \hat{r}_j]$ is normalized to be random variable ρ_j in the range $[-1, 1]$. We aim to see how well different approaches can model the uncertainties of different ρ_js while eliminating potential effects of different r_js and $\hat{r}_j$s on the performance. For it, we set the variables $r_j \equiv 1.02$ and $\hat{r}_j \equiv 0.2$ for all $j \in J$, and let the distributions of ρ_js to be different from each other. As a result, the distributions of different returns have the same center and same half width, but different peakedness. Another reason for this setup

is that we can compare different portfolio solutions just by comparing their true probability of shortfall because all solutions have the same true expected return, i.e., 1.02; otherwise, the probability of shortfall and expected returns can both be different for different solutions, and comparing different solutions would usually require introducing the preference function, which unnecessarily complicates our discussion. We consider $J = 10$ portfolios and the ρ_js are distributed as follows:

$$\text{prob}(\rho_j = x) = \frac{c_j}{2 - 2 \cdot e^{-c_j}} e^{-c_j|x|}, \ -1 \leq x \leq 1,$$

where $c_j = 0.1 \cdot (J + 1 - j)^2$. Because we know the true distribution of the return for each portfolio, we can calculate the true probability of shortfall and true expected return for a given portfolio solution.

In this section, we set $s = 1.0$ and thus the shortfall probability becomes the probability of loss. We first generate $N = 100$ samples of portfolio data. We assume that the decision maker first chooses an in-sample expected return value, then finds the parameters for the RO models and the MIP model that achieve this specified in-sample expected return. Then we calculate and compare the true probability of shortfall of the solutions from different models. We repeat this procedure for 10 different values of in-sample expected return that are evenly spaced in the widest possible range. Figure 4.4(a) shows the results of true shortfall probability for cases with different in-sample expected return values for the case $N = 100$. The results for $N = 300$ and $N = 3000$ are shown in Figure 4.4(b) and Figure 4.4(c), respectively. Note that we do not have the result for the MIP problem when $N = 3000$ because we were not able to solve the MIP problem optimally.

We first compare the solutions from RO models with different uncertainty sets. From Figure 4.4, we see that except the first case in Figure 4.4(b) and the first case in Figure 4.4(c), the performance of the order statistic uncertainty set is better than or at least the same as other uncertainty

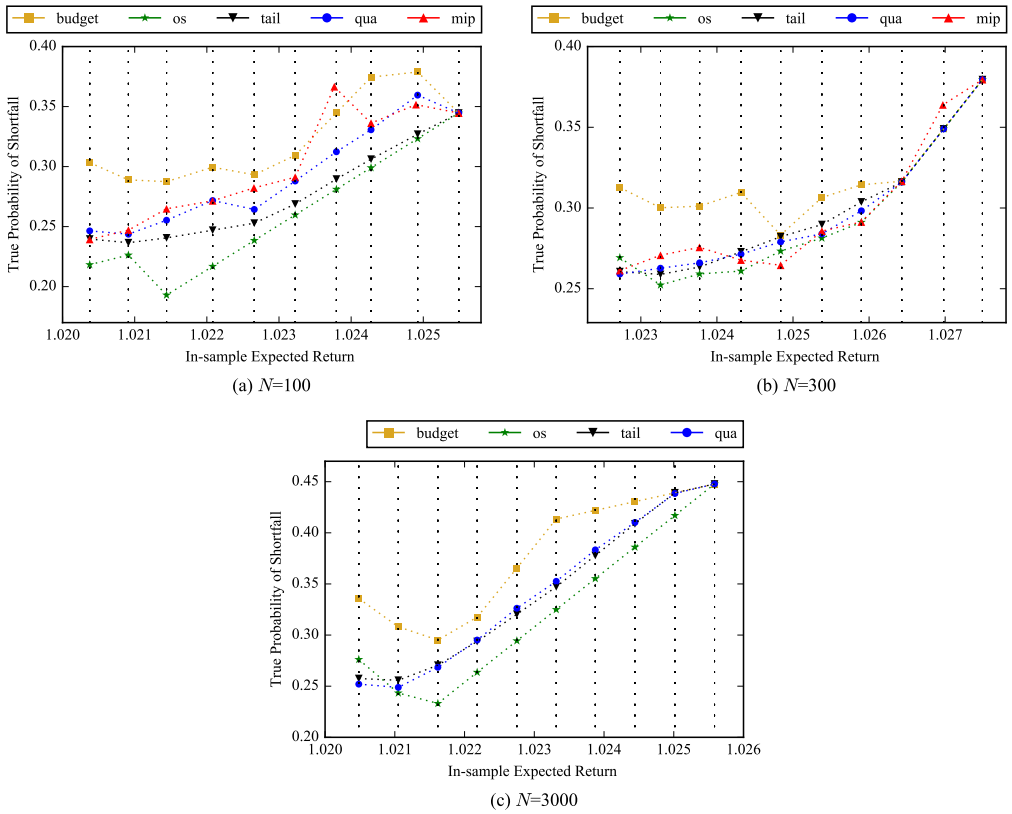


Figure 4.4: True shortfall probability versus the in-sample expected return for cases of $N = 100, 300, 3000$; the true expected return is always 1.02.

sets because the solutions of RO model with the order statistic uncertainty set have lower or the same true probability of shortfall than other uncertainty sets. The superior performance of the order statistic uncertainty set can be explained by its three advantages illustrated in Section 4.3.3.

Another observation for all three cases in the Figure 4.4 is that when the in-sample expected return is the largest, the solutions of all models have the same true probability of shortfall. The reason is that in such cases, all uncertainty sets are almost identical because they are all very small and collapse to the point $\mathbf{Z} = \mathbf{0}$.

Next, we are interested to see whether the RO model with the order statistic uncertainty set can provide better results than the MIP model. For the 20 cases in Figure 4.4(a) and Figure 4.4(b),

the solution of the RO model with the order statistic uncertainty set is no worse than that of the MIP formulation, except for two cases in Figure 4.4(b) with smallest and fifth smallest in-sample expected return. The performance of the RO model with the order statistic uncertainty set is better than the MIP formulation in most cases.

4.5.2 Experiments with Real Data

We test the RO models on the standard financial datasets with daily returns from Fama-French website. We choose two categories, each of which contains datasets with different numbers of portfolios. In each category, we choose three data files: one with a small number of portfolios, one with a large number of portfolios and one with a median number of portfolios. Summary statistics of the 6 data files are in Table 4.2. Each data file has a value-weighted portfolio dataset and an equally-weighted (EW) portfolio dataset; we have 12 datasets in total. For each dataset, we use the data with date no later than July 31 2018, and the begin date in each data file can be found in Table 4.2. There are 11903 dates in data file with ID=3 and 490 dates with ID=6 that have missing data, so we remove those dates from corresponding data files; other data files do not have missing data. Table 4.2 shows the number of remaining observations for each dataset after removing the dates with missing data. Of all the portfolios in each equally-weighted dataset, we find the portfolio with the minimum average return and report its average return and the standard deviation (SD) of its return, which can be found in Table 4.2 along with the maximum average portfolio return and associated standard deviation. Each dataset is equally split into two groups, and we use the first 50% of observations as the training set, and the remaining 50% as the test set. We acknowledge the above setup may not be perfectly realistic; for example, 50/50% split of data does not reflect the rolling horizon procedure. However, we believe the potential bias in our setup does not affect the relative performance of different methods.

Table 4.2: Selected Fama-French data files

ID	Data files	Begin date[1]	# of observations[2]	Min average portfolio return and associated SD (EW)	Max average portfolio return and associated SD (EW)
1	10 Industry Portfolios [Daily]	07-01-1926	24286	1.00062 / 0.011920	1.00098 / 0.011290
2	30 Industry Portfolios [Daily]	07-01-1926	24286	1.00062 / 0.011920	1.00112 / 0.023939
3	49 Industry Portfolios [Daily]	07-01-1969	12383	1.00055 / 0.007460	1.00104 / 0.011161
4	6 Portfolios Formed on Size and Long-Term Reversal (2×3) [Daily]	03-20-1930	23185	1.00044 / 0.011566	1.00147 / 0.012531
5	10 Portfolios Formed on Long-Term Reversal [Daily]	03-20-1930	23185	1.00052 / 0.012303	1.00168 / 0.013931
6	25 Portfolios Formed on Size and Long-Term Reversal (5 x 5) [Daily]	03-20-1930	22695	1.00040 / 0.011993	1.00187 / 0.013289

[1] All the end dates are 07-31-2018.

[2] This column shows the number of observations after excluding the missing data.

We use 0.99 for the threshold return s in the shortfall constraint rather than 1.0. The reason we do not use $s = 1.0$ is that the expected return of each portfolio is very close to 1.0, so if we choose $s = 1.0$, then the range of the shortfall probability is very narrow. If we use $s = 0.99$, we would have a much wider range of shortfall probability.

We conduct experiments for RO models with the order statistic uncertainty set, the budget uncertainty set and the ellipsoidal uncertainty set. We do not have results for the MIP model and the RO model with the tail uncertainty set because the results for these models could not be obtained in a reasonable amount of time. Note that the tail uncertainty set introduces the same number of decision variables as the number of observations, which is over 20000 for most of our datasets and the computation becomes difficult.

For each dataset, we solve each model with different parameters such that the corresponding solutions have 20 different values of in-sample expected return, which are equally spaced between the lowest possible value $\underline{r}$ and the highest possible value $\bar{r}$ of the in-sample expected return. Consequently, for each dataset, each model would have 20 different solutions that have in-sample expected return as $r_m = \underline{r} + (\bar{r} - \underline{r}) \cdot m/19, m = 0, 1, \cdots, 19$. The values of $\underline{r}$ and $\bar{r}$ depend on the dataset.

We compare different uncertainty sets in a pair-wise fashion. For a dataset, we compare the solutions of two uncertainty sets $\mathscr{U}_1$ and $\mathscr{U}_2$ using the out-of-sample expected return and out-of-sample shortfall probability. We define a counter function $d(\cdot)$ for each uncertainty set, which has the initial value 0. For each in-sample expected return r_m, we update the counter function as follows: if the solution of $\mathscr{U}_1$ (or $\mathscr{U}_2$) has a higher out-of-sample average return and lower out-of-sample probability of shortfall than the solution of $\mathscr{U}_2$ (or $\mathscr{U}_1$, respectively), then we increase $d(\mathscr{U}_1)$ $\big($or $d(\mathscr{U}_2)$, respectively$\big)$ by 1. For 20 cases of each dataset, the value $d(\mathscr{U}_1)$ is essentially the number of cases that the solution of $\mathscr{U}_1$ has better out-of-sample performance than $\mathscr{U}_2$. For some cases, the solutions of the two uncertainty sets are not comparable, e.g., if the solution of $\mathscr{U}_1$ has a higher out-of-sample average return and higher out-of-sample probability of shortfall than the solution of $\mathscr{U}_2$, then both $d(\mathscr{U}_1)$ and $d(\mathscr{U}_2)$ will not change. Therefore, for each dataset, we have $d(\mathscr{U}_1)+d(\mathscr{U}_2)\leq 20$.

Figure 4.5(a) is the result of comparison between the budget uncertainty set and the order statistic uncertainty set. Each point shows the values of $d(\mathscr{U}^{OS})$ and $d(\mathscr{U}^{B})$ for a dataset. We see that $d(\mathscr{U}^{OS})$ is greater than $d(\mathscr{U}^{B})$ for 8 datasets (the points in the shaded area). In contrast, $d(\mathscr{U}^{B})$ is greater than $d(\mathscr{U}^{OS})$ only for 3 datasets, and in one dataset $d(\mathscr{U}^{OS})=d(\mathscr{U}^{B})$. This means that out of 12 instances that could be ordered, in 8 instances $\mathscr{U}^{OS}$ outperforms $\mathscr{U}^{B}$.

Figure 4.5(b) shows the comparison between the ellipsoidal uncertainty set and the order statistic uncertainty set. The results can be interpreted in a similar fashion. We see that $d(\mathscr{U}^{OS})$ is greater than $d(\mathscr{U}^{Q})$ for 4 datasets (the points in the shaded area), and $d(\mathscr{U}^{Q})$ is greater than $d(\mathscr{U}^{OS})$ for 7 datasets. The ellipsoidal uncertainty set is thus better than the order statistic uncertainty set, which is in contrast with the result in Section 4.5.1. The discrepancy can be explained by the arguments in Section 4.3.3. The distributions of Z_js in Section 4.5.1 are different from each other, and as explained in Section 4.3.3, the order statistic uncertainty set can capture richer infor-

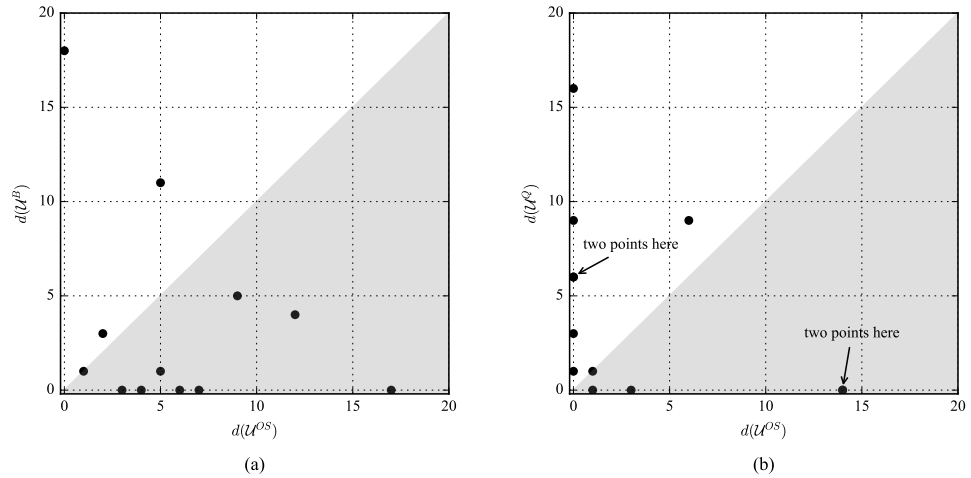


Figure 4.5: Comparison of budget uncertainty set and order statistic uncertainty set (left); comparison of ellipsoidal uncertainty set and order statistic uncertainty set (right).

mation about the heterogeneity of different distributions. In contrast, the distributions of different Z_js in this section are very similar, so the quadratic form of the ellipsoidal uncertainty set is a better fit for the data due to its nonlinear structure.

4.6 Conclusion

In this paper, we developed the order statistic uncertainty set for robust linear optimization models. Data-free and distribution-free properties of random variables are embedded in the design of the order statistic uncertainty set. High geometric flexibility of the order statistic uncertainty set enables it to capture richer information of distributions than existing uncertainty sets. We analyzed the advantages of the order statistic uncertainty set, and showed that three existing uncertainty sets are special cases of the new uncertainty set. Numerical experiments on portfolio optimization problems with shortfall constraints showed that the robust optimization model with the order statistic uncertainty set has better performance than the robust model with other uncertainty sets,

as well as the sample average approximation for the stochastic optimization model.

Appendices

Appendix A

Appendix for Chapter 2

A.1 Proof of Proposition 1.

Proof. Proof of Proposition 1 Because $\boldsymbol{x} = \mathbf{0}$ is feasible to the problem (2.4), the problem (2.4) has at least one solution. Now suppose the optimal solutions to the problem (2.4) is not unique. Assume any two different optimal solutions $\boldsymbol{x}'$ and $\boldsymbol{x}''$. If the optimal objective value is 0, then we must have $\boldsymbol{x}' = \boldsymbol{x}'' = \mathbf{0}$, and so the solution is unique. Now suppose the optimal objective value is strictly larger than zero, and then none of $\boldsymbol{x}'$ and $\boldsymbol{x}''$ can be the $\mathbf{0}$ vector.

Because $\boldsymbol{x}'$ is not $\mathbf{0}$, we must be able to find a i such that $x'_i > 0$. Next we prove $x'_k > 0, \forall k = 1, \cdots, n$. Suppose not, then we must be able to find a k such that $x'_k = 0$. Because the network is strongly connected, we must be able to find a directed path from i to k as follows: $i \to j_1 \to \cdots \to j_{m-1} \to j_m \to k$ such that $\alpha_{i,j_1} > 0,\ \alpha_{j_m,k} > 0,\ \alpha_{j_l,j_{l+1}} > 0,\ \forall l = 1, \cdots, m-1$. Then we have the following

$$
\begin{aligned}
x'_k &\geq x'_{j_m} \cdot \alpha_{j_m,k} \\
&\geq x'_{j_{m-1}} \cdot \alpha_{j_{m-1},j_m} \cdot \alpha_{j_m,k} \\
&\geq x'_{j_{m-2}} \cdot \alpha_{j_{m-2},j_{m-1}} \cdot \alpha_{j_{m-1},j_m} \cdot \alpha_{j_m,k} \\
&\vdots \\
&\geq x'_i \cdot \alpha_{i,j_1} \cdots \alpha_{j_{m-2},j_{m-1}} \cdot \alpha_{j_{m-1},j_m} \cdot \alpha_{j_m,k} \\
&> 0,
\end{aligned}
$$

which contradicts $x'_k = 0$. Therefore, we have proven that $x'_k > 0, \forall k = 1, \cdots, n$. Similarly, we have $x''_k > 0, \forall k = 1, \cdots, n$.

Denote $i_1 = \arg\min_i \frac{x''_i}{x'_i}$ and denote $t_{i_1} = t_{min} = \frac{x''_{i_1}}{x'_{i_1}}$. Next we prove that there exists $t > 0$ such that $\boldsymbol{x}'' = t \cdot \boldsymbol{x}'$ holds. Assume it is not true, then we must be able to find $i_2 \neq i_1$, and $t_{i_2} = \frac{x''_{i_1}}{x'_{i_1}} > t_{i_1}$. We then construct a new vector $\overline{\boldsymbol{x}} = \boldsymbol{x}'' - t_{min} \cdot \boldsymbol{x}'$. According to the definition of t_{min}, we must have $\overline{x}_{i_1} = 0$, $\overline{x}_{i_2} > 0$ and $\overline{x}_i \geq 0, \forall i \neq i_1, i_2$. Because the network is strongly connected, we must be able to find a directed path from i_2 to i_1 as follows: $i_2 \to k_1 \to \cdots \to k_{r-1} \to k_r \to i_1$ such that $\alpha_{i_2,k_1} > 0,\ \alpha_{k_r,i_1} > 0,\ \alpha_{k_l,k_{l+1}} > 0,\ \forall l = 1, \cdots, r-1$. Because $\boldsymbol{x}', \boldsymbol{x}''$ both satisfy (2.4b) and the equalities in (2.4b) are linear, we have

$$\overline{x}_j = \sum_{i=1}^{n} \overline{x}_i \cdot \alpha_{ij}, \forall j = 1, \cdots, n$$

We apply the above relation to the nodes along the path $i_2 \to k_1 \to \cdots \to k_{r-1} \to k_r \to i_1$ and we have

$$\begin{aligned}
\overline{x}_{i_1} &\geq \overline{x}_{k_r} \cdot \alpha_{k_r,i_1} \\
&\geq \overline{x}_{k_{r-1}} \cdot \alpha_{k_{r-1},k_r} \cdot \alpha_{k_r,i_1} \\
&\geq \overline{x}_{k_{r-2}} \cdot \alpha_{k_{r-2},k_{r-1}} \cdot \alpha_{k_{r-1},k_r} \cdot \alpha_{k_r,i_1} \\
&\vdots \\
&\geq \overline{x}_{i_2} \cdot \alpha_{i_2,k_1} \cdots \alpha_{k_{r-2},k_{r-1}} \cdot \alpha_{k_{r-1},k_r} \cdot \alpha_{k_r,i_1} \\
&> 0,
\end{aligned}$$

which contradicts $\overline{x}_{i_1} = 0$. Therefore, there must exist $t > 0$ such that $\boldsymbol{x}'' = t \cdot \boldsymbol{x}'$ holds.

Because $\boldsymbol{x}'$ and $\boldsymbol{x}''$ are two different solutions, $t \neq 1$. Considering $\boldsymbol{x}'' = t \cdot \boldsymbol{x}'$, we then know

that the objective values for the solutions $\boldsymbol{x}'$ and $\boldsymbol{x}''$ are not equal to each other. This contradicts the fact that $\boldsymbol{x}'$ and $\boldsymbol{x}''$ are both optimal solutions. Therefore, we have proven the uniqueness of $\boldsymbol{x}^*$. □

A.2 Proof of Proposition 2.

Part 1: we prove $\hat{r}(t)$ (weakly) monotonically increases.

We first show that $\hat{r}(t) \leq \min\{\eta, 1\}$. We discuss two cases. Case 1: $\eta \leq 1$. In this case, if $\hat{r}(t) > \min\{\eta, 1\} = \eta$, then there are $\sum_i X_i(t) = \sum_{j=1}^n \left(r_j(t) \cdot x_j^*\right) \geq \sum_{j=1}^n \left(\hat{r}(t) \cdot x_j^*\right) > \eta \cdot \sum_{j=1}^n x_j^*$ vehicles in the network. This cannot happen because the total number of vehicles is more than $\eta \cdot \sum_{j=1}^n x_j^*$. Therefore, $\hat{r}(t) > \min\{\eta, 1\}$ cannot happen in this case. Case 2: $\eta > 1$. In this case the ratio $r_j(t) = \frac{X_j(t)}{x_j^*}$ for the critical location is bounded by 1, so we have $\hat{r}(t) \leq 1 = \min\{\eta, 1\}$. Therefore, we have proven $\hat{r}(t) \leq \min\{\eta, 1\}$.

Define the unused inventory at location j during time period t: $UI_j(t) = I_j(t-1) - X_j(t)$. Then we know $UI_j(t) \geq 0, \forall j = 1, \cdots, n, \forall t$. For any $j = 1, \cdots, n$ and $t \geq 2$, we have

$$
\begin{aligned}
X_j(t) &= \min\left\{\sum_{k=1}^n Q_{jk},\ I_j(t-1)\right\} \\
&= \min\left\{\sum_{k=1}^n Q_{jk},\ I_j(t-2) - X_j(t-1) + I_j(t-1) - I_j(t-2) + X_j(t-1)\right\} \\
&= \min\left\{\sum_{k=1}^n Q_{jk},\ UI_j(t-1) + I_j(t-1) - I_j(t-2) + X_j(t-1)\right\} \\
&\overset{(1.A)}{=} \min\left\{\sum_{k=1}^n Q_{jk},\ UI_j(t-1) + \sum_{i=1}^n Y_{ij}(t-1)\right\} \\
&\overset{(1.B)}{=} \min\left\{\sum_{k=1}^n Q_{jk},\ UI_j(t-1) + \sum_{i=1}^n X_i(t-1) \cdot \alpha_{ij}\right\}
\end{aligned}
$$

where the equality $(1.A)$ is due to the relation (2.3) and the equality $(1.B)$ is due to the relation (2.2). We then prove that $\hat{r}(t)$ is non-decreasing in time t. It suffices to prove that $X_j(t+1) \geq \hat{r}(t) \cdot x_j^*, \forall j = 1, \cdots, n, \forall t \geq 1$. We have

$$\begin{aligned}
X_j(t+1) &= \min\left\{ \sum_{k=1}^{n} Q_{jk},\ UI_j(t) + \sum_{i=1}^{n} X_i(t) \cdot \alpha_{ij} \right\} \\
&\geq \min\left\{ \sum_{k=1}^{n} Q_{jk},\ \sum_{i=1}^{n} X_i(t) \cdot \alpha_{ij} \right\} \\
&\geq \min\left\{ x_j^*,\ \sum_{i=1}^{n} X_i(t) \cdot \alpha_{ij} \right\} \\
&= \min\left\{ x_j^*,\ \sum_{i=1}^{n} r_i(t) \cdot x_i^* \cdot \alpha_{ij} \right\} \\
&\geq \min\left\{ x_j^*,\ \sum_{i=1}^{n} \hat{r}(t) \cdot x_i^* \cdot \alpha_{ij} \right\} \\
&= \min\left\{ x_j^*,\ \hat{r}(t) \cdot x_j^* \right\} \\
&= \hat{r}(t) \cdot x_j^*
\end{aligned}$$

where the last equality is due to $\hat{r}(t) \leq \min\{\eta, 1\} \leq 1$.

Part 2: we prove $\min\{\eta, 1\} - \hat{r}(t+2) \leq C_r \cdot \left[\min\{\eta, 1\} - \hat{r}(t)\right]$**.**

Let $\Delta\hat{r}(t) = \min\{\eta, 1\} - \hat{r}(t)$. For time $t \geq 1$, there must exists a location, say, location j_1, such that $I_{j_1}(t-1) \geq \eta \cdot x_{j_1}^*$ because otherwise the total vehicles in the network would be strictly less than $\eta \cdot \sum_{j=1}^{n} x_j^*$ at the end of time $t-1$.

For any fixed location j_3, we have

$$X_{j_3}(t+2)$$

$$
\begin{aligned}
&= \min\left\{ \sum_{j=1}^{n} Q_{j_3,j},\ UI_{j_3}(t+1) + \sum_{j_2=1}^{n} X_{j_2}(t+1)\cdot\alpha_{j_2,j_3} \right\} \\
&\geq \min\left\{ x^*_{j_3},\ \sum_{j_2=1}^{n} X_{j_2}(t+1)\cdot\alpha_{j_2,j_3} \right\} \\
&\geq \min\left\{ x^*_{j_3},\ \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\min\left[\sum_{j=1}^{n} Q_{j_2,j},\ UI_{j_2}(t) + \sum_{j=1}^{n} X_j(t)\cdot\alpha_{j,j_2} \right] \right\} \\
&\geq \min\left\{ x^*_{j_3},\ \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\min\left[x^*_{j_2},\ \sum_{j=1}^{n} X_j(t)\cdot\alpha_{j,j_2} \right] \right\} \\
&= \min\left\{ x^*_{j_3},\ \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\min\left[x^*_{j_2},\ X_{j_1}(t)\cdot\alpha_{j_1,j_2} + \sum_{j\neq j_1} X_j(t)\cdot\alpha_{j,j_2} \right] \right\} \\
&\geq \min\left\{ x^*_{j_3},\ \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\min\left[x^*_{j_2},\ \min\left\{ \sum_{j} Q_{j_1,j},\ I_{j_1}(t-1) \right\}\cdot\alpha_{j_1,j_2} + \sum_{j\neq j_1} \hat{r}(t)\cdot x^*_j\cdot\alpha_{j,j_2} \right] \right\} \\
&\geq \min\left\{ x^*_{j_3},\ \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\min\left[x^*_{j_2},\ \min\left\{ x^*_{j_1},\ \eta\cdot x^*_{j_1} \right\}\cdot\alpha_{j_1,j_2} + \sum_{j\neq j_1} \hat{r}(t)\cdot x^*_j\cdot\alpha_{j,j_2} \right] \right\} \\
&= \min\left\{ x^*_{j_3},\ \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\min\left[x^*_{j_2},\ \Delta\hat{r}(t)\cdot x^*_{j_1}\cdot\alpha_{j_1,j_2} + \sum_{j=1}^{n} \hat{r}(t)\cdot x^*_j\cdot\alpha_{j,j_2} \right] \right\} \\
&= \min\left\{ x^*_{j_3},\ \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\min\left[x^*_{j_2},\ \Delta\hat{r}(t)\cdot x^*_{j_1}\cdot\alpha_{j_1,j_2} + \hat{r}(t)\cdot x^*_{j_2} \right] \right\} \\
&= \min\left\{ x^*_{j_3},\ \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\hat{r}(t)\cdot x^*_{j_2} + \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\min\left[x^*_{j_2} - \hat{r}(t)\cdot x^*_{j_2},\ \Delta\hat{r}(t)\cdot x^*_{j_1}\cdot\alpha_{j_1,j_2} \right] \right\} \\
&= \min\left\{ x^*_{j_3},\ \hat{r}(t)\cdot x^*_{j_3} + \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\min\left[x^*_{j_2} - \hat{r}(t)\cdot x^*_{j_2},\ \Delta\hat{r}(t)\cdot x^*_{j_1}\cdot\alpha_{j_1,j_2} \right] \right\} \\
&\geq \min\left\{ x^*_{j_3},\ \hat{r}(t)\cdot x^*_{j_3} + \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\min\left[\Delta\hat{r}(t)\cdot x^*_{j_2},\ \Delta\hat{r}(t)\cdot x^*_{j_1}\cdot\alpha_{j_1,j_2} \right] \right\} \\
&\overset{(2.A)}{=} \min\left\{ x^*_{j_3},\ \hat{r}(t)\cdot x^*_{j_3} + \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\Delta\hat{r}(t)\cdot x^*_{j_1}\cdot\alpha_{j_1,j_2} \right\} \\
&\overset{(2.B)}{=} \hat{r}(t)\cdot x^*_{j_3} + \sum_{j_2=1}^{n} \alpha_{j_2,j_3}\cdot\Delta\hat{r}(t)\cdot x^*_{j_1}\cdot\alpha_{j_1,j_2}
\end{aligned}
$$

where the equality $(2.A)$ is due to the following relation: $\Delta\hat{r}(t)\cdot x^*_{j_2} = \Delta\hat{r}(t)\cdot\sum_j\left(x^*_j\cdot\alpha_{j,j_2}\right) > \Delta\hat{r}(t)\cdot x^*_{j_1}\cdot\alpha_{j_1,j_2}$. The relation $(2.B)$ is due to the following relation:

$$
\begin{aligned}
& x^*_{j_3} \\
\geq & \hat{r}(t)\cdot x^*_{j_3} + \Delta\hat{r}(t)\cdot x^*_{j_3} \\
= & \hat{r}(t)\cdot x^*_{j_3} + \Delta\hat{r}(t)\cdot\sum_{j_2=1}^{n} x^*_{j_2}\cdot\alpha_{j_2,j_3} \\
= & \hat{r}(t)\cdot x^*_{j_3} + \Delta\hat{r}(t)\cdot\sum_{j_2=1}^{n}\left[\left(\sum_j x^*_j\cdot\alpha_{j,j_2}\right)\cdot\alpha_{j_2,j_3}\right] \\
> & \hat{r}(t)\cdot x^*_{j_3} + \Delta\hat{r}(t)\cdot\sum_{j_2=1}^{n}\left[\left(x^*_{j_1}\cdot\alpha_{j_1,j_2}\right)\cdot\alpha_{j_2,j_3}\right]
\end{aligned}
$$

Then we have

$$
\begin{aligned}
& \hat{r}(t+2) - \hat{r}(t) \\
= & \min_{j_3\in\{1,\cdots,n\}} r^{j_3}_{t+2} - \hat{r}(t) \\
= & \min_{j_3\in\{1,\cdots,n\}}\left\{\frac{X_{j_3}(t+2)}{x^*_{j_3}}\right\} - \hat{r}(t) \\
\geq & \min_{j_1,j_3\in\{1,\cdots,n\}}\left\{\frac{\hat{r}(t)\cdot x^*_{j_3} + \sum_{j_2=1}^{n}\alpha_{j_2,j_3}\cdot\Delta\hat{r}(t)\cdot x^*_{j_1}\cdot\alpha_{j_1,j_2}}{x^*_{j_3}}\right\} - \hat{r}(t) \\
= & \hat{r}(t) + \min_{j_1,j_3\in\{1,\cdots,n\}}\left\{\frac{\sum_{j_2=1}^{n}\alpha_{j_2,j_3}\cdot\Delta\hat{r}(t)\cdot x^*_{j_1}\cdot\alpha_{j_1,j_2}}{x^*_{j_3}}\right\} - \hat{r}(t) \\
= & \Delta\hat{r}(t)\cdot\min_{j_1,j_3\in\{1,\cdots,n\}}\left\{\frac{x^*_{j_1}}{x^*_{j_3}}\cdot\sum_{j_2=1}^{n}\alpha_{j_1,j_2}\cdot\alpha_{j_2,j_3}\right\}
\end{aligned}
$$

Then we have

$$\Delta\hat{r}(t+2)$$

$$
\begin{aligned}
=&[\min\{\eta,1\}-\hat{r}(t+2)]-\Delta\hat{r}(t)+\Delta\hat{r}(t)\\
=&[\min\{\eta,1\}-\hat{r}(t+2)]-[\min\{\eta,1\}-\hat{r}(t)]+\Delta\hat{r}(t)\\
=&\hat{r}(t)-\hat{r}(t+2)+\Delta\hat{r}(t)\\
\leq&\Delta\hat{r}(t)-\Delta\hat{r}(t)\cdot\min_{j_1,j_3\in\{1,\cdots,n\}}\left\{\frac{x^*_{j_1}}{x^*_{j_3}}\cdot\sum_{j_2=1}^{n}\alpha_{j_1,j_2}\cdot\alpha_{j_2,j_3}\right\}\\
=&\Delta\hat{r}(t)\cdot\left[1-\min_{j_1,j_3\in\{1,\cdots,n\}}\left\{\frac{x^*_{j_1}}{x^*_{j_3}}\cdot\sum_{j_2=1}^{n}\alpha_{j_1,j_2}\cdot\alpha_{j_2,j_3}\right\}\right]\\
=&C_r\cdot\Delta\hat{r}(t)
\end{aligned}
$$

Because the network is complete and has at least three locations, we know that $\sum_{j_2=1}^{n}\alpha_{j_1,j_2}\cdot\alpha_{j_2,j_3}>0,\forall j_1,j_3=1,\cdots,n$. Therefore, we have $C_r<1$.

Part 3: we prove $\lim_{t\to\infty}\hat{r}(t)=\min\{\eta,1\}$**.**

As we have defined, $\Delta\hat{r}(t)=\min\{\eta,1\}-\hat{r}(t)$. Because we have proven that $\hat{r}(t)\leq\min\{\eta,1\}$, we have $\Delta\hat{r}(t)\geq 0$. Considering $\Delta\hat{r}(t+2)\leq C_r\cdot\Delta\hat{r}(t)$ and $C_r<1$, we must have $\lim_{t\to\infty}\Delta\hat{r}(t)=0$, or equivalently $\lim_{t\to\infty}\hat{r}(t)=\min\{\eta,1\}$.

Part 4: we prove the globally stable equilibrium, i.e., $\lim_{t\to\infty}X_j(t)=\min\{\eta,1\}\cdot x^*_j,\forall j=1,\cdots,n$**.**

Part 4.1: we prove for the case that $\eta\leq 1$. According to the definition of $\hat{r}(t)$, we have $r_k(t)\geq\hat{r}(t),\forall k=1,\cdots,n$. So $\lim_{t\to\infty}r_k(t)\geq\lim_{t\to\infty}\hat{r}(t)=\min\{\eta,1\}=\eta,\forall k=1,\cdots,n$.

We also have that $\sum_{j=1}^{n}X_j(t)=\sum_{j=1}^{n}r_j(t)\cdot x^*_j\leq\eta\cdot\sum_{j=1}^{n}x^*_j$. So

$$
\begin{aligned}
&\lim_{t\to\infty}r_k(t)\\
=&\lim_{t\to\infty}\frac{1}{x^*_k}\cdot\left[\eta\cdot\sum_{j=1}^{n}x^*_j-\sum_{j\neq k}r_j(t)\cdot x^*_j\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{1}{x_k^*} \cdot \left[\eta \cdot \sum_{j=1}^{n} x_j^* - \sum_{j \neq k} \eta \cdot x_j^* \right] \\
&= \frac{1}{x_k^*} \cdot \eta \cdot x_k^* \\
&= \eta
\end{aligned}
$$

So we have $\eta \leq \lim_{t\to\infty} r_k(t) \leq \eta, \forall k = 1, \cdots, n$, which leads to $\lim_{t\to\infty} r_k(t) = \eta, \forall k = 1, \cdots, n$.

Part 4.2: we prove for the case that $\eta > 1$.

In this case, we have $\lim_{t\to\infty} \hat{r}(t) = \min\{\eta, 1\} = 1$. We first prove for the critical location(s). Assume node l_1 is a critical location and then we have $r_{l_1}(t) = \frac{X_{l_1}(t)}{x_{l_1}^*} \leq \frac{\sum_{k=1}^{n} Q_{l_1,k}}{x_{l_1}^*} = 1$. So $\lim_{t\to\infty} r_{l_1}(t) \leq 1$. Because we have $\lim_{t\to\infty} \hat{r}(t) = 1$, we have $\lim_{t\to\infty} r_{l_1}(t) \geq \lim_{t\to\infty} \hat{r}(t) = 1$. Therefore, we have proven $1 \leq \lim_{t\to\infty} r_{l_1}(t) \leq 1$. Thus we have $\lim_{t\to\infty} r_{l_1}(t) = 1$, or equivalently $\lim_{t\to\infty} X_{l_1}(t) = \min\{\eta, 1\} \cdot x_{l_1}^* = x_{l_1}^*$.

Next we prove that if node l_2 is a non-critical location, then $\lim_{t\to\infty} X_{l_2}(t) = \min\{\eta, 1\} \cdot x_{l_2}^* = x_{l_2}^*$ holds. Denote the set of all non-critical location as S_N. We have $x_j^* < \sum_{k=1}^{n} Q_{jk}, \forall j \in S_N$. According to Lemma 1 (stated and proved later), we have $\lim_{t\to\infty} \sum_{j\in S_N} I_j(t) = \sum_{j\in S_N} x_j^*$. If node l_2 is a non-critical location, then

$$
\begin{aligned}
&\lim_{t\to\infty} X_{l_2}(t) \\
&= \lim_{t\to\infty} \left(\sum_{j\in S_N} X_j(t) - \sum_{j\neq l_2, j\in S_N} X_j(t) \right) \\
&= \lim_{t\to\infty} \sum_{j\in S_N} X_j(t) - \lim_{t\to\infty} \sum_{j\neq l_2, j\in S_N} X_j(t) \\
&\leq \lim_{t\to\infty} \sum_{j\in S_N} I_j(t-1) - \lim_{t\to\infty} \sum_{j\neq l_2, j\in S_N} \hat{r}(t) \cdot x_j^* \\
&\leq \sum_{j\in S_N} x_j^* - \sum_{j\neq l_2, j\in S_N} x_j^*
\end{aligned}
$$

$$= x^*_{l_2}$$

We also have $\lim_{t\to\infty} X_{l_2}(t) \geq \lim_{t\to\infty} \hat{r}(t) \cdot x^*_{l_2} = x^*_{l_2}$. So we have $x^*_{l_2} \leq \lim_{t\to\infty} X_{l_2}(t) \leq x^*_{l_2}$. We then have $\lim_{t\to\infty} X_{l_2}(t) = x^*_{l_2}$.

Therefore, we have proven that if $\eta > 1$, then $\lim_{t\to\infty} X_j(t) = x^*_j, \forall j = 1, \cdots, n$.

Lemma 1. *If* $\eta \geq 1$, *then* $\lim_{t\to\infty} \sum_{j\in S_N} I_j(t) = \sum_{j\in S_N} x^*_j$ *holds.*

Proof. Proof of Lemma 1

We prove it by showing that both $\lim_{t\to\infty} \sum_{j\in S_N} I_j(t) \geq \sum_{j\in S_N} x^*_j$ and $\lim_{t\to\infty} \sum_{j\in S_N} I_j(t) \leq \sum_{j\in S_N} x^*_j$ holds.

We first prove that $\lim_{t\to\infty} \sum_{j\in S_N} I_j(t) \geq \sum_{j\in S_N} x^*_j$. Because $X_j(t) \geq \hat{r}(t) \cdot x^*_j, \forall j$ holds, we have that $\lim_{t\to\infty} \sum_{j\in S_N} I_j(t) \geq \lim_{t\to\infty} \sum_{j\in S_N} X_j(t+1) \geq \lim_{t\to\infty} \sum_{j\in S_N} \hat{r}(t+1) \cdot x^*_j = \lim_{t\to\infty} \hat{r}(t+1) \cdot \sum_{j\in S_N} x^*_j = \sum_{j\in S_N} x^*_j$

We then prove that $\lim_{t\to\infty} \sum_{j\in S_N} I_j(t) \leq \sum_{j\in S_N} x^*_j$. We prove it by showing the following: for any $\varepsilon_N > 0$, there exists a t_N such that $\sum_{j\in S_N} I_j(t) - \sum_{j\in S_N} x^*_j \leq \varepsilon_N, \forall t \geq t_N$. We assume that ε_N is no greater than $\min_{j\in S_N}(\sum_i Q_{ji} - x^*_j)$, which is positive according to the definition of S_N; this means $\varepsilon_N \leq \min_{j\in S_N}(\sum_i Q_{ji} - x^*_j)$. Let $\varepsilon_1 = \min_{ij}\{\alpha_{ij}\} \cdot \varepsilon_N \cdot \frac{1}{4\cdot n^2}$. Because $\lim_{t\to\infty} X_j(t) \geq \lim_{t\to\infty} \hat{r}(t) \cdot x^*_j = x^*_j, \forall j = 1, \cdots, n$, we can find a t_1 such that $X_j(t) \geq x^*_j - \varepsilon_1, \forall t \geq t_1, \forall j = 1, \cdots, n$.

For $t \geq t_1$, we have

$$\sum_{j\in S_N} I_j(t) \tag{A.1a}$$

$$= \sum_{j\in S_N} I_j(t-1) - \sum_{j\in S_N} X_j(t) + \sum_{j\in S_N} \sum_{i=1}^{n} Y_{ij}(t) \tag{A.1b}$$

$$= \sum_{j\in S_N} I_j(t-1) - \sum_{j\in S_N} \sum_{i=1}^{n} Y_{ji}(t) + \sum_{j\in S_N} \sum_{i=1}^{n} Y_{ij}(t) \tag{A.1c}$$

$$= \sum_{j\in S_N} I_j(t-1) - \sum_{j\in S_N}\sum_{i\in S_N} Y_{ji}(t) - \sum_{j\in S_N}\sum_{i\notin S_N} Y_{ji}(t) + \sum_{j\in S_N}\sum_{i\in S_N} Y_{ij}(t) + \sum_{j\in S_N}\sum_{i\notin S_N} Y_{ij}(t) \tag{A.1d}$$

$$= \sum_{j\in S_N} I_j(t-1) - \sum_{j\in S_N}\sum_{i\notin S_N} Y_{ji}(t) + \sum_{j\in S_N}\sum_{i\notin S_N} Y_{ij}(t) \tag{A.1e}$$

$$= \sum_{j\in S_N} I_j(t-1) - \sum_{j\in S_N}\sum_{i\notin S_N} X_j(t)\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N} X_i(t)\cdot\alpha_{ij} \tag{A.1f}$$

$$\leq \sum_{j\in S_N} I_j(t-1) - \sum_{j\in S_N}\sum_{i\notin S_N} X_j(t)\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N}\sum_{k} Q_{ik}\cdot\alpha_{ij} \tag{A.1g}$$

$$= \sum_{j\in S_N} I_j(t-1) - \sum_{j\in S_N}\sum_{i\notin S_N} X_j(t)\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N} x_i^*\cdot\alpha_{ij} \tag{A.1h}$$

Next we study two results.

Result A: For time $t \geq t_1$, if $\sum_{j\in S_N} I_j(t) - \sum_{j\in S_N} x_j^* \leq \frac{1}{2}\cdot\varepsilon_N$ holds, then $\sum_{j\in S_N} I_j(t+1) - \sum_{j\in S_N} x_j^* < \varepsilon_N$.

We prove Result A. Suppose $t \geq t_1$ and $\sum_{j\in S_N} I_j(t) - \sum_{j\in S_N} x_j^* \leq \frac{1}{2}\cdot\varepsilon_N$. Based on (A.1h), we have

$$\sum_{j\in S_N} I_j(t+1) \tag{A.2a}$$

$$\leq \sum_{j\in S_N} I_j(t) - \sum_{j\in S_N}\sum_{i\notin S_N} X_j(t+1)\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N} x_i^*\cdot\alpha_{ij} \tag{A.2b}$$

$$\leq \sum_{j\in S_N} I_j(t) - \sum_{j\in S_N}\sum_{i\notin S_N} (x_j^* - \varepsilon_1)\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N} x_i^*\cdot\alpha_{ij} \tag{A.2c}$$

$$= \sum_{j\in S_N} I_j(t) + \sum_{j\in S_N}\sum_{i\notin S_N} \varepsilon_1\cdot\alpha_{ji} - \sum_{j\in S_N}\sum_{i\notin S_N} x_j^*\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N} x_i^*\cdot\alpha_{ij} \tag{A.2d}$$

$$= \sum_{j\in S_N} I_j(t) + \sum_{j\in S_N}\sum_{i\notin S_N} \varepsilon_1\cdot\alpha_{ji} - \sum_{j\in S_N}\sum_{i\notin S_N} x_j^*\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N} x_i^*\cdot\alpha_{ij} - \sum_{j\in S_N}\sum_{i\in S_N} x_j^*\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\in S_N} x_i^*\cdot\alpha_{ij} \tag{A.2e}$$

$$= \sum_{j\in S_N} I_j(t) + \sum_{j\in S_N}\sum_{i\notin S_N} \varepsilon_1 \cdot \alpha_{ji} - \sum_{j\in S_N}\sum_{i=1}^{n} x_j^* \cdot \alpha_{ji} + \sum_{j\in S_N}\sum_{i=1}^{n} x_i^* \cdot \alpha_{ij} \tag{A.2f}$$

$$= \sum_{j\in S_N} I_j(t) + \sum_{j\in S_N}\sum_{i\notin S_N} \varepsilon_1 \cdot \alpha_{ji} - \sum_{j\in S_N} x_j^* + \sum_{j\in S_N} x_j^* \tag{A.2g}$$

$$= \sum_{j\in S_N} I_j(t) + \sum_{j\in S_N}\sum_{i\notin S_N} \varepsilon_1 \cdot \alpha_{ji} \tag{A.2h}$$

$$< \frac{1}{2}\cdot\varepsilon_N + \sum_{j\in S_N} x_j^* + \sum_{j\in S_N}\sum_{i=1}^{n} \varepsilon_1 \cdot \alpha_{ji} \tag{A.2i}$$

$$= \frac{1}{2}\cdot\varepsilon_N + \sum_{j\in S_N} x_j^* + \sum_{j\in S_N} \varepsilon_1 \tag{A.2j}$$

$$< \frac{1}{2}\cdot\varepsilon_N + \sum_{j\in S_N} x_j^* + n\cdot\varepsilon_1 \tag{A.2k}$$

$$= \frac{1}{2}\cdot\varepsilon_N + \sum_{j\in S_N} x_j^* + n\cdot\min_{ij}\{\alpha_{ij}\}\cdot\varepsilon_N\cdot\frac{1}{4\cdot n^2} \tag{A.2l}$$

$$= \frac{1}{2}\cdot\varepsilon_N + \sum_{j\in S_N} x_j^* + \min_{ij}\{\alpha_{ij}\}\cdot\varepsilon_N\cdot\frac{1}{4\cdot n} \tag{A.2m}$$

$$< \varepsilon_N + \sum_{j\in S_N} x_j^* \tag{A.2n}$$

Therefore, we have proven Result A.

Result B: For time $t \geq t_1$, if $\sum_{j\in S_N} I_j(t) - \sum_{j\in S_N} x_j^* \geq \frac{1}{2}\cdot\varepsilon_N$ holds, then $\sum_{j\in S_N} I_j(t+1) - \sum_{j\in S_N} I_j(t) \leq -\frac{1}{4\cdot n}\cdot\varepsilon_N\cdot\min_{ij}\{\alpha_{ij}\}$.

We prove Result B. Denote the number of nodes in S_N as $|S_N|$. If $\sum_{j\in S_N} I_j(t) - \sum_{j\in S_N} x_j^* \geq \frac{1}{2}\cdot\varepsilon_N$, then we must have a node $j_0 \in S_N$ such that $I_{j_0}(t) - x_{j_0}^* \geq \frac{1}{2\cdot n}\cdot\varepsilon_N$ because otherwise $\sum_{j\in S_N} I_j(t) - \sum_{j\in S_N} x_j^* < |S_N|\cdot\frac{1}{2\cdot n}\cdot\varepsilon_N < n\cdot\frac{1}{2\cdot n}\cdot\varepsilon_N = \frac{1}{2}\cdot\varepsilon_N$. Then based on (A.1h), we have

$$\sum_{j\in S_N} I_j(t+1) \tag{A.3a}$$

$$\leq \sum_{j\in S_N} I_j(t) - \sum_{j\in S_N}\sum_{i\notin S_N} X_j(t+1)\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N} x_i^*\cdot\alpha_{ij} \tag{A.3b}$$

$$= \sum_{j\in S_N} I_j(t) - \sum_{i\notin S_N} X_{j_0}(t+1)\cdot\alpha_{j_0,i} - \sum_{j\in S_N\setminus\{j_0\}} \sum_{i\notin S_N} X_j(t+1)\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N} x_i^*\cdot\alpha_{ij} \tag{A.3c}$$

$$\overset{(4.A)}{\leq} \sum_{j\in S_N} I_j(t) - \sum_{i\notin S_N}\left(x_{j_0}^* + \frac{1}{2\cdot n}\cdot\varepsilon_N\right)\cdot\alpha_{j_0,i} - \sum_{j\in S_N\setminus\{j_0\}} \sum_{i\notin S_N} (x_j^* - \varepsilon_1)\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N} x_i^*\cdot\alpha_{ij}$$

$$= \sum_{j\in S_N} I_j(t) - \sum_{i\notin S_N}\frac{1}{2\cdot n}\cdot\varepsilon_N\cdot\alpha_{j_0,i} + \sum_{j\in S_N\setminus\{j_0\}} \sum_{i\notin S_N} \varepsilon_1\cdot\alpha_{ji} - \sum_{j\in S_N}\sum_{i\notin S_N} x_j^*\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N} x_i^*\cdot\alpha_{ij} \tag{A.3d}$$

$$= \sum_{j\in S_N} I_j(t) - \sum_{i\notin S_N}\frac{1}{2\cdot n}\cdot\varepsilon_N\cdot\alpha_{j_0,i} + \sum_{j\in S_N\setminus\{j_0\}} \sum_{i\notin S_N} \varepsilon_1\cdot\alpha_{ji} - \sum_{j\in S_N}\sum_{i\in S_N} x_j^*\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\in S_N} x_i^*\cdot\alpha_{ij} - \sum_{j\in S_N}\sum_{i\notin S_N} x_j^*\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i\notin S_N} x_i^*\cdot\alpha_{ij} \tag{A.3e}$$

$$= \sum_{j\in S_N} I_j(t) - \sum_{i\notin S_N}\frac{1}{2\cdot n}\cdot\varepsilon_N\cdot\alpha_{j_0,i} + \sum_{j\in S_N\setminus\{j_0\}} \sum_{i\notin S_N} \varepsilon_1\cdot\alpha_{ji} - \sum_{j\in S_N}\sum_{i=1}^{n} x_j^*\cdot\alpha_{ji} + \sum_{j\in S_N}\sum_{i=1}^{n} x_i^*\cdot\alpha_{ij} \tag{A.3f}$$

$$= \sum_{j\in S_N} I_j(t) - \sum_{i\notin S_N}\frac{1}{2\cdot n}\cdot\varepsilon_N\cdot\alpha_{j_0,i} + \sum_{j\in S_N\setminus\{j_0\}} \sum_{i\notin S_N} \varepsilon_1\cdot\alpha_{ji} - \sum_{j\in S_N} x_j^* + \sum_{j\in S_N} x_j^* \tag{A.3g}$$

$$= \sum_{j\in S_N} I_j(t) - \sum_{i\notin S_N}\frac{1}{2\cdot n}\cdot\varepsilon_N\cdot\alpha_{j_0,i} + \sum_{j\in S_N\setminus\{j_0\}} \sum_{i\notin S_N} \varepsilon_1\cdot\alpha_{ji} \tag{A.3h}$$

$$\leq \sum_{j\in S_N} I_j(t) - \frac{1}{2\cdot n}\cdot\varepsilon_N\cdot\min_{ij}\{\alpha_{ij}\} + \sum_{j\in S_N\setminus\{j_0\}} \varepsilon_1 \tag{A.3i}$$

$$\leq \sum_{j\in S_N} I_j(t) - \frac{1}{2\cdot n}\cdot\varepsilon_N\cdot\min_{ij}\{\alpha_{ij}\} + n\cdot\varepsilon_1 \tag{A.3j}$$

$$\leq \sum_{j\in S_N} I_j(t) - \frac{1}{2\cdot n}\cdot\varepsilon_N\cdot\min_{ij}\{\alpha_{ij}\} + n\cdot\min_{ij}\{\alpha_{ij}\}\cdot\varepsilon_N\cdot\frac{1}{4\cdot n^2} \tag{A.3k}$$

$$\leq \sum_{j\in S_N} I_j(t) - \frac{1}{4\cdot n}\cdot\varepsilon_N\cdot\min_{ij}\{\alpha_{ij}\} \tag{A.3l}$$

where for the inequality $(4.A)$, we have used $X_{j_0}(t+1) = \min\left\{\sum_{k=1}^{n} Q_{j_0,k},\ I_{j_0}(t)\right\} \geq \min\left\{x^*_{j_0} + \varepsilon_N,\ x^*_{j_0} + \frac{1}{2\cdot n}\cdot\varepsilon_N\right\} = x^*_{j_0} + \frac{1}{2\cdot n}\cdot\varepsilon_N$. Therefore, we have proven Result B.

According to Result B, if $\sum_{j\in S_N} I_j(t_1) \geq \sum_{j\in S_N} x^*_j + \varepsilon_N$, then $\sum_{j\in S_N} I_j(t_1)$ will decrease by at least $\frac{1}{4\cdot n}\cdot\varepsilon_N\cdot\min_{ij}\{\alpha_{ij}\}$, which is a positive constant. Then there must be a time $t_2 \geq t_1$ such that $\sum_{j\in S_N} I_j(t_2) \leq \sum_{j\in S_N} x^*_j + \varepsilon_N$. We next show that $\forall t \geq t_2$, we have $\sum_{j\in S_N} I_j(t) \leq \sum_{j\in S_N} x^*_j + \varepsilon_N$. It suffices to prove that $\forall t \geq t_2$, if $\sum_{j\in S_N} I_j(t) \leq \sum_{j\in S_N} x^*_j + \varepsilon_N$, then $\sum_{j\in S_N} I_j(t+1) \leq \sum_{j\in S_N} x^*_j + \varepsilon_N$. We discuss two cases. (1) The first case: $\sum_{j\in S_N} I_j(t) \leq \sum_{j\in S_N} x^*_j + \frac{1}{2}\cdot\varepsilon_N$. According to Result A, we have $\sum_{j\in S_N} I_j(t+1) \leq \sum_{j\in S_N} x^*_j + \varepsilon_N$. (2) For the second case, $\sum_{j\in S_N} x^*_j + \frac{1}{2}\cdot\varepsilon_N \leq \sum_{j\in S_N} I_j(t) \leq \sum_{j\in S_N} x^*_j + \varepsilon_N$. According to Result B, we have $\sum_{j\in S_N} I_j(t+1) \leq \sum_{j\in S_N} I_j(t) - \frac{1}{4\cdot n}\cdot\varepsilon_N\cdot\min_{ij}\{\alpha_{ij}\} < \sum_{j\in S_N} x^*_j + \varepsilon_N$. Therefore, we have proven that $\sum_{j\in S_N} I_j(t) \leq \sum_{j\in S_N} x^*_j + \varepsilon_N, \forall t \geq t_2$. So we have proven $\lim_{t\to\infty}\sum_{j\in S_N} I_j(t) \leq \sum_{j\in S_N} x^*_j$. This completes the proof for the lemma.

□

Part 5: we prove the draining effect.

If $\eta \geq 1$, then $\lim_{t\to\infty} I_j(t) \geq \lim_{t\to\infty} X_j(t) = x^*_j, \forall j \in S_N$. We also have $\lim_{t\to\infty}\sum_{j\in S_N} I_j(t) = \sum_{j\in S_N} x^*_j$. If $k \in S_N$, then

$$\begin{aligned}
&\lim_{t\to\infty} I_k(t)\\
=&\lim_{t\to\infty}\sum_{j\in S_N} I_j(t) - \lim_{t\to\infty}\sum_{j\neq k, j\in S_N} I_j(t)\\
\leq&\sum_{j\in S_N} x^*_j - \lim_{t\to\infty}\sum_{j\neq k, j\in S_N} x^*_j\\
=&x^*_k.
\end{aligned}$$

So we have that $x^*_j \leq \lim_{t\to\infty} I_j(t) \leq x^*_j, \forall j \in S_N$. Therefore we have $\lim_{t\to\infty} I_j(t) = x^*_j, \forall j \in S_N$. Hence no idle vehicles will be at any non-critical locations. □

A.3 Proof of Proposition 3.

Note that in this proof, we will use Proposition 6, which is also proven in the Appendix.

When the outbound demands of the location j are $\sum_{k\neq j} Q_{jk}, \forall k \neq j$, denote the equilibrium flow as $\boldsymbol{x}^*$. Then we have $x_j^* = \sum_{k=1}^n Q_{jk}$ and $x_v^* < \sum_{k=1}^n Q_{vk}, \forall v \neq j$.

Assume an infinitesimal quantity $\tau > 0$. When the outbound demands of the location j become $Q_{jk} + \tau \cdot \varepsilon_{jk}, \forall k \neq j$, denote the equilibrium flow as $\widetilde{\boldsymbol{x}}^*$. According to Proposition 6, the location j remains to be the only critical location as long as τ is small enough, which will be assumed throughout this proof. Then we have $\widetilde{x}_j^* = \sum_{k=1}^n Q_{jk}$ and $\widetilde{x}_v^* < \sum_{k=1}^n Q_{vk}, \forall v \neq j$.

Because we have

$$
\begin{aligned}
x_j^* &= \sum_{i\neq j} x_i^* \cdot \alpha_{ij}, \\
\widetilde{x}_j^* &= \sum_{i\neq j} \widetilde{x}_i^* \cdot \alpha_{ij},
\end{aligned}
$$

we then have

$$
\begin{aligned}
& \sum_{i\neq j} (\widetilde{x}_i^* - x_i^*) \cdot \alpha_{ij} = \widetilde{x}_j^* - x_j^* = \tau, \\
\Longrightarrow \quad & \sum_{i\neq j} (\widetilde{x}_i^* - x_i^*) \cdot \min_k \alpha_{kj} \leq \tau \leq \sum_{i\neq j} (\widetilde{x}_i^* - x_i^*) \cdot \max_k \alpha_{kj} \\
\Longrightarrow \quad & \frac{\tau}{\max_k \alpha_{kj}} \leq \sum_{i\neq j} (\widetilde{x}_i^* - x_i^*) \leq \frac{\tau}{\min_k \alpha_{kj}}.
\end{aligned}
$$

Then we have

$$
\sigma(\tau, \epsilon_j) = \widetilde{x}_j^* - x_j^* + \sum_{i\neq j} (\widetilde{x}_i^* - x_i^*)
$$

$$
\begin{aligned}
&\geq \tau + \frac{\tau}{\max_k \alpha_{kj}} \\
&= \tau \cdot \left(1 + \frac{1}{\max_k \alpha_{kj}}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\sigma(\tau, \epsilon_j) =& \widetilde{x}_j^* - x_j^* + \sum_{i \neq j} (\widetilde{x}_i^* - x_i^*) \\
&\leq \tau + \frac{\tau}{\min_k \alpha_{kj}} \\
&= \tau \cdot \left(1 + \frac{1}{\min_k \alpha_{kj}}\right)
\end{aligned}
$$

Therefore, we have proven that $\lambda_j(\epsilon_j) = \lim_{\tau \to 0} \frac{\sigma(\tau, \epsilon_j)}{\tau} \geq \left(1 + \frac{1}{\max_k \alpha_{kj}}\right)$ and $\lambda_j(\epsilon_j) \leq \left(1 + \frac{1}{\min_k \alpha_{kj}}\right)$.

□

A.4 Proof of Proposition 4.

Note that in this proof, we will use Proposition 6 and Proposition 7, which are both proven in the Appendix.

Assume a positive infinitesimal small quantity τ. We consider the following three cases.

1. Scenario 1: at the beginning of each period, there are Q_{mv} potential customers seeking to go from location m to location v. At the beginning of $t = 1$, we place $\sum_{m=1}^{n} Q_{vm}$ vehicles at each location $v \in \{1, \cdots, n\}$. Denote the flow from location m to location v during the time period t as $Y_{mv}(t)$. Denote the outbound flow from location v during time period t as $X_v(t)$. We use $UI_v(t)$ to denote the number of vehicles that are not used by any customer during the period t and stay at location v during the period t. We use TF to denote the total network flow in the equilibrium.

2. Scenario 2: at the beginning of each period, there are Q_{vm} potential customers seeking to go from location v to location m (except from location i to location k), and $Q_{ik}+\tau$ potential customers seeking to go from location i to location k. At the beginning of $t=1$, we place $\sum_{v=1}^{n} Q_{mv}$ vehicles at each location $v \in \{1,\cdots,n\} \setminus \{i\}$, and $\sum_{m=1}^{n} Q_{im}+\tau$ vehicles at the location i. Denote the flow from location m to location v during the time period t as $Y'_{mv}(t)$. Denote the outbound flow from location v during time period t as $X'_v(t)$. We use $UI'_v(t)$ to denote the number of vehicles that are not used by any customer during the period t and stay at location v during the period t. We use TF' to denote the total network flow in the equilibrium.

3. Scenario 3: at the beginning of each period, there are Q_{vm} potential customers seeking to go from location v to location m (except from location j to location k), and $Q_{jk}+\tau$ potential customers seeking to go from location j to location k. At the beginning of $t=1$, we place $\sum_{v=1}^{n} Q_{mv}$ vehicles at each location $v \in \{1,\cdots,n\} \setminus \{j\}$, and $\sum_{m=1}^{n} Q_{jm}+\tau$ vehicles at the location j. Denote the flow from location m to location v during the time period t as $Y''_{mv}(t)$. Denote the outbound flow from location v during time period t as $X''_v(t)$. We use $UI''_v(t)$ to denote the number of vehicles that are not used by any customer during the period t and stay at location v during the period t. We use TF'' to denote the total network flow in the equilibrium.

In Scenario 1, for each arc (m,v), denote the associated fraction $\frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}}$ as α_{mv}.

In Scenario 2, for each arc (m,v) such that $m \neq i$, denote the associated fraction $\frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}}$ as α'_{mv}; for the arc (i,m) such that $m \neq k$, denote the associated fraction $\frac{Q_{im}}{\tau+\sum_{l=1}^{n} Q_{il}}$ as α'_{im}; for the arc (i,k), denote the associated fraction $\frac{Q_{ik}+\tau}{\tau+\sum_{l=1}^{n} Q_{il}}$ as α'_{ik}.

In Scenario 3, for each arc (m,v) such that $m \neq j$, denote the associated fraction $\frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}}$ as α''_{mv}; for the arc (j,m) such that $m \neq k$, denote the associated fraction $\frac{Q_{jm}}{\tau+\sum_{l=1}^{n} Q_{jl}}$ as α''_{jm}; for the arc

(j,k), denote the associated fraction $\frac{Q_{jk}+\tau}{\tau+\sum_{l=1}^{n} Q_{jl}}$ as α''_{jk}.

According to Proposition 2, we know the system state in each of the above 3 scenarios will evolve into the corresponding equilibrium flow. Denote the equilibrium flow in Scenario 1, Scenario 2 and Scenario 3 as $\boldsymbol{x}^*$, $\widetilde{\boldsymbol{x}}^*$ and $\widehat{\boldsymbol{x}}^*$. According to Lemma 2, we know $X_v(t) \geq x_v^*$, $X'_v(t) \geq \widetilde{x}_v^*$, and $X''_v(t) \geq \widehat{x}_v^*$ $\forall v = 1, \cdots, n,\ \forall t \geq 1$.

Because τ is infinitesimal small, we know that the location j is the only critical location in Scenario 2 according to Proposition 7, and the location j is the only critical location in Scenario 3 according to Proposition 6.

Part 1: We show that the following relation holds.

$$\tau + X'_j(t) = X''_j(t),\ \forall t \geq 2, \tag{A.4a}$$

$$\frac{X'_i(t)}{\tau + \sum_{m=1}^{n} Q_{im}} \leq \frac{X''_i(t)}{\sum_{m=1}^{n} Q_{im}},\ \forall t \geq 2, \tag{A.4b}$$

$$X'_v(t) \leq X''_v(t),\ \forall v \neq i, j,\ \forall t \geq 2, \tag{A.4c}$$

$$UI'_v(t) \leq UI''_v(t),\ \forall v \neq j,\ \forall t \geq 2 \tag{A.4d}$$

Because the location j is the critical location in Scenario 2 and Scenario 3, we must have $X'_j(t) = \sum_{l=1}^{n} Q_{jl},\ \forall t \geq 1$ and $X''_j(t) = \tau + \sum_{l=1}^{n} Q_{jl},\ \forall t \geq 1$. So the first equality in (A.4) holds for all $t \geq 2$. Then we just need to prove (A.4b), (A.4c) and (A.4d).

First we prove that (A.4b), (A.4c) and (A.4d) hold for $t = 2$. For the first period $t = 1$, we have

$$X'_v(1) = \sum_{l=1}^{n} Q_{vl}, \forall v \neq i$$

$$X'_i(1) = \tau + \sum_{l=1}^{n} Q_{il},$$

$$UI'_v(1) = 0,\ \forall v \neq j,$$

$$X_v''(1) = \sum_{l=1}^{n} Q_{vl}, \forall v \neq j$$

$$X_j''(1) = \tau + \sum_{l=1}^{n} Q_{jl},$$

$$UI_v''(1) = 0,\ \forall v \neq j.$$

Then for all $v \neq i, j, k$, we have

$$\begin{aligned} UI_v'(2) &= \max\left\{0,\ UI_v'(1) + \sum_m Y_{mv}'(1) - \sum_m Q_{vm}\right\} \\ &= \max\left\{0,\ 0 + X_i'(1)\cdot\frac{Q_{iv}}{\tau + \sum_l Q_{il}} + \sum_{m\neq i} X_m'(1)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} - \sum_m Q_{vm}\right\} \\ &= \max\left\{0,\ \sum_m Q_{mv} - \sum_m Q_{vm}\right\} \end{aligned}$$

and

$$\begin{aligned} UI_v''(2) &= \max\left\{0,\ UI_v''(1) + \sum_m Y_{mv}''(1) - \sum_m Q_{vm}\right\} \\ &= \max\left\{0,\ 0 + X_j''(1)\cdot\frac{Q_{jv}}{\tau + \sum_l Q_{jl}} + \sum_{m\neq j} X_m''(1)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} - \sum_m Q_{vm}\right\} \\ &= \max\left\{0,\ \sum_m Q_{mv} - \sum_m Q_{vm}\right\}. \end{aligned}$$

So we have $UI_v'(2) \leq UI_v''(2),\ v \neq i, j, k.$

For the location i, we have

$$UI_i'(2) = \max\left\{0,\ UI_i'(1) + \sum_m Y_{mi}'(1) - \tau - \sum_m Q_{im}\right\}$$

$$= \max\left\{0,\ 0 + \sum_m X'_m(1) \cdot \frac{Q_{mi}}{\sum_l Q_{ml}} - \tau - \sum_m Q_{im}\right\}$$
$$= \max\left\{0,\ \sum_m Q_{mi} - \tau - \sum_m Q_{im}\right\}$$

and

$$\begin{aligned} UI''_i(2) &= \max\left\{0,\ UI''_i(1) + \sum_m Y''_{mi}(1) - \sum_m Q_{im}\right\} \\ &= \max\left\{0,\ 0 + X''_j(1) \cdot \frac{Q_{ji}}{\tau + \sum_l Q_{jl}} + \sum_{m \neq j} X''_m(1) \cdot \frac{Q_{mi}}{\sum_l Q_{ml}} - \sum_m Q_{im}\right\} \\ &= \max\left\{0,\ Q_{ji} + \sum_{m \neq j} Q_{mi} - \sum_m Q_{im}\right\} \\ &= \max\left\{0,\ \sum_m Q_{mi} - \sum_m Q_{im}\right\} \end{aligned}$$

So we have $UI'_i(2) \leq UI''_i(2)$.

For the location k, we have

$$\begin{aligned} UI'_k(2) &= \max\left\{0,\ UI'_k(1) + \sum_m Y'_{mk}(1) - \sum_m Q_{km}\right\} \\ &= \max\left\{0,\ 0 + X'_i(1) \cdot \frac{\tau + Q_{ik}}{\tau + \sum_l Q_{il}} + \sum_{m \neq i} X'_m(1) \cdot \frac{Q_{mk}}{\sum_l Q_{ml}} - \sum_m Q_{km}\right\} \\ &= \max\left\{0,\ \tau + \sum_m Q_{mk} - \sum_m Q_{km}\right\} \end{aligned}$$

and

$$UI''_k(2) = \max\left\{0,\ UI''_k(1) + \sum_m Y''_{mk}(1) - \sum_m Q_{km}\right\}$$

$$= \max\left\{0,\ 0 + X_j''(1)\cdot\frac{\tau+Q_{jk}}{\tau+\sum_l Q_{jl}} + \sum_{m\neq j} X_m''(1)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} - \sum_m Q_{km}\right\}$$
$$= \max\left\{0,\ \tau + \sum_m Q_{mk} - \sum_m Q_{km}\right\}.$$

So we have $UI_k'(2) \leq UI_k''(2)$.

Therefore we have proven that $UI_v'(2) \leq UI_v''(2),\ v \neq j$.

For location $v \neq i, j, k$, we have

$$X_v'(2) = \min\left\{\sum_m Q_{vm},\ UI_v'(1) + \sum_m Y_{mv}'(1)\right\}$$
$$= \min\left\{\sum_m Q_{vm},\ X_i'(1)\cdot\frac{Q_{iv}}{\tau+\sum_l Q_{il}} + X_j'(1)\cdot\frac{Q_{jv}}{\sum_l Q_{jl}} + \sum_{m\neq i,j} X_m'(1)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}}\right\}$$

and

$$X_v''(2) = \min\left\{\sum_m Q_{vm},\ UI_v''(1) + \sum_m Y_{mv}''(1)\right\}$$
$$= \min\left\{\sum_m Q_{vm},\ X_j''(1)\cdot\frac{Q_{jv}}{\tau+\sum_l Q_{jl}} + \sum_{m\neq j} X_m''(1)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}}\right\}$$
$$= \min\left\{\sum_m Q_{vm},\ \left(\tau + \sum_l Q_{jl}\right)\cdot\frac{Q_{jv}}{\tau+\sum_l Q_{jl}} + \sum_{m\neq j} X_m''(1)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}}\right\}$$
$$= \min\left\{\sum_m Q_{vm},\ Q_{jv} + X_i''(1)\cdot\frac{Q_{iv}}{\sum_l Q_{il}} + \sum_{m\neq i,j} X_m''(1)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}}\right\}.$$

Because we have

$$
\begin{aligned}
&X'_j(1)\cdot\frac{Q_{jv}}{\sum_l Q_{jl}} \le \big(\sum_l Q_{jl}\big)\cdot\frac{Q_{jv}}{\sum_l Q_{jl}} = Q_{jv},\\
&X'_i(1)\cdot\frac{Q_{iv}}{\tau+\sum_l Q_{il}} = Q_{iv} = X''_i(1)\cdot\frac{Q_{iv}}{\sum_l Q_{il}},\\
&X'_m(1) = \sum_l Q_{ml} = X''_m(1), \forall m \neq i, j,
\end{aligned}
$$

we have

$$
\begin{aligned}
&X'_v(2)\\
=&\min\left\{\sum_m Q_{vm},\ UI'_v(1)+X'_i(1)\cdot\frac{Q_{iv}}{\tau+\sum_l Q_{il}}+X'_j(1)\cdot\frac{Q_{jv}}{\sum_l Q_{jl}}+\sum_{m\neq i,j}X'_m(1)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}}\right\}\\
\le&\min\left\{\sum_m Q_{vm},\ Q_{jv}+X''_i(1)\cdot\frac{Q_{iv}}{\sum_l Q_{il}}+\sum_{m\neq i,j}X''_m(1)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}}\right\}\\
=&X''_v(2).
\end{aligned}
$$

For the location i, we have

$$
\begin{aligned}
X'_i(2) =&\min\left\{\tau+\sum_m Q_{im},\ UI'_i(1)+\sum_m Y'_{mi}(1)\right\}\\
=&\min\left\{\tau+\sum_m Q_{im},\ X'_j(1)\cdot\frac{Q_{ji}}{\sum_l Q_{jl}}+\sum_{m\neq i,j}X'_m(1)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}}\right\}\\
=&\min\left\{\tau+\sum_m Q_{im},\ \sum_l Q_{jl}\cdot\frac{Q_{ji}}{\sum_l Q_{jl}}+\sum_{m\neq i,j}\sum_l Q_{ml}\cdot\frac{Q_{mi}}{\sum_l Q_{ml}}\right\}\\
=&\min\left\{\tau+\sum_m Q_{im},\ Q_{ji}+\sum_{m\neq i,j}Q_{mi}\right\}
\end{aligned}
$$

and

$$
\begin{aligned}
X_i''(2) &= \min\left\{ \sum_m Q_{im},\ UI_i''(1) + \sum_{m\neq i} Y_{mi}''(1) \right\} \\
&= \min\left\{ \sum_m Q_{im},\ X_j''(1)\cdot\frac{Q_{ji}}{\tau+\sum_l Q_{jl}} + \sum_{m\neq i,j} X_m''(1)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{im},\ \left(\tau+\sum_l Q_{jl}\right)\cdot\frac{Q_{ji}}{\tau+\sum_l Q_{jl}} + \sum_{m\neq i,j}\sum_l Q_{ml}\cdot\frac{Q_{mi}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{im},\ Q_{ji} + \sum_{m\neq i,j} Q_{mi} \right\}.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\frac{X_i'(2)}{\tau+\sum_{m=1}^n Q_{jm}} \\
=&\frac{\min\left\{ \tau+\sum_m Q_{im},\ Q_{ji}+\sum_{m\neq i,j} Q_{mi} \right\}}{\tau+\sum_{m=1}^n Q_{im}} \\
=&\min\left\{ \frac{\tau+\sum_m Q_{im}}{\tau+\sum_{m=1}^n Q_{im}},\ \frac{Q_{ji}+\sum_{m\neq i,j} Q_{mi}}{\tau+\sum_{m=1}^n Q_{im}} \right\} \\
\leq&\min\left\{ \frac{\sum_m Q_{im}}{\sum_{m=1}^n Q_{im}},\ \frac{Q_{ji}+\sum_{m\neq i,j} Q_{mi}}{\sum_{m=1}^n Q_{im}} \right\} \\
=&\frac{\min\left\{ \sum_m Q_{im},\ Q_{ji}+\sum_{m\neq i,j} Q_{mi} \right\}}{\sum_{m=1}^n Q_{im}} \\
=&\frac{X_i''(2)}{\sum_{m=1}^n Q_{im}}.
\end{aligned}
$$

So $\frac{X_i'(2)}{\tau+\sum_{m=1}^n Q_{im}} \leq \frac{X_i''(2)}{\sum_{m=1}^n Q_{im}}$ is proven.

For the location k, we have

$$
\begin{aligned}
X'_k(2) &= \min\left\{ \sum_m Q_{km},\ UI'_k(1) + \sum_m Y'_{mk}(1) \right\} \\
&= \min\left\{ \sum_m Q_{km},\ X'_i(1)\cdot\frac{\tau+Q_{ik}}{\tau+\sum_l Q_{il}} + X'_j(1)\cdot\frac{Q_{jk}}{\sum_l Q_{jl}} + \sum_{m\neq i,j} X'_m(1)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{km},\ \left(\tau+\sum_l Q_{il}\right)\cdot\frac{\tau+Q_{ik}}{\tau+\sum_l Q_{il}} + \left(\sum_l Q_{jl}\right)\cdot\frac{Q_{jk}}{\sum_l Q_{jl}} \right. \\
&\qquad\qquad \left. + \sum_{m\neq i,j}\left(\sum_l Q_{ml}\right)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{km},\ \tau + Q_{ik} + Q_{jk} + \sum_{m\neq i,j} Q_{mk} \right\}
\end{aligned}
$$

and

$$
\begin{aligned}
X''_k(2) &= \min\left\{ \sum_m Q_{km},\ UI''_k(1) + \sum_m Y''_{mk}(1) \right\} \\
&= \min\left\{ \sum_m Q_{km},\ X''_i(1)\cdot\frac{Q_{ik}}{\sum_l Q_{il}} + X''_j(1)\cdot\frac{\tau+Q_{jk}}{\tau+\sum_l Q_{jl}} + \sum_{m\neq i,j} X''_m(1)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{km},\ X''_i(1)\cdot\frac{Q_{ik}}{\sum_l Q_{il}} + X''_j(1)\cdot\frac{\tau+Q_{jk}}{\tau+\sum_l Q_{jl}} + \sum_{m\neq i,j} X''_m(1)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{km},\ \left(\sum_l Q_{il}\right)\cdot\frac{Q_{ik}}{\sum_l Q_{il}} + \left(\tau+\sum_l Q_{jl}\right)\cdot\frac{\tau+Q_{jk}}{\tau+\sum_l Q_{jl}} \right. \\
&\qquad\qquad \left. + \sum_{m\neq i,j}\left(\sum_l Q_{ml}\right)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{km},\ Q_{ik} + \tau + Q_{jk} + \sum_{m\neq i,j} Q_{mk} \right\}.
\end{aligned}
$$

So we have $X'_k(2) \leq X''_k(2)$.

Therefore, we have proven that the relation (A.4b), (A.4c) and (A.4d) hold for the time period

$t = 2$. Suppose the relation (A.4b), (A.4c) and (A.4d) hold for the time period $t \geq 2$, we next show that they also hold for the time period $t+1$.

For all $v \neq i, j, k$, we have

$$\begin{aligned}
UI'_v(t+1) &= \max\left\{0,\ UI'_v(t) + \sum_m Y'_{mv}(t) - \sum_m Q_{vm}\right\} \\
&= \max\left\{0,\ UI'_v(t) + X'_i(t)\cdot\frac{Q_{iv}}{\tau+\sum_l Q_{il}} + X'_j(t)\cdot\frac{Q_{jv}}{\sum_l Q_{jl}}\right. \\
&\qquad\left. + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} - \sum_m Q_{vm}\right\} \\
&= \max\left\{0,\ UI'_v(t) + X'_i(t)\cdot\frac{Q_{iv}}{\tau+\sum_l Q_{il}} + \left(\sum_l Q_{jl}\right)\cdot\frac{Q_{jv}}{\sum_l Q_{jl}}\right. \\
&\qquad\left. + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} - \sum_m Q_{vm}\right\} \\
&= \max\left\{0,\ UI'_v(t) + X'_i(t)\cdot\frac{Q_{iv}}{\tau+\sum_l Q_{il}} + Q_{jv}\right. \\
&\qquad\left. + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} - \sum_m Q_{vm}\right\}
\end{aligned}$$

and

$$\begin{aligned}
UI''_v(t+1) &= \max\left\{0,\ UI''_v(t) + \sum_m Y''_{mv}(t) - \sum_m Q_{vm}\right\} \\
&= \max\left\{0,\ UI''_v(t) + X''_i(t)\cdot\frac{Q_{iv}}{\sum_l Q_{il}} + X''_j(t)\cdot\frac{Q_{jv}}{\tau+\sum_l Q_{jl}}\right. \\
&\qquad\left. + \sum_{m\neq i,j} X''_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} - \sum_m Q_{vm}\right\} \\
&= \max\left\{0,\ UI''_v(t) + X''_i(t)\cdot\frac{Q_{iv}}{\sum_l Q_{il}} + \left(\tau+\sum_l Q_{jl}\right)\cdot\frac{Q_{jv}}{\tau+\sum_l Q_{jl}}\right.
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad + \sum_{m\neq i,j} X''_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} - \sum_m Q_{vm}\Bigg\} \\
&= \max\Bigg\{0,\ UI''_v(t) + X''_i(t)\cdot\frac{Q_{iv}}{\sum_l Q_{il}} + Q_{jv} \\
&\qquad\qquad + \sum_{m\neq i,j} X''_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} - \sum_m Q_{vm}\Bigg\} \\
&\geq \max\Bigg\{0,\ UI'_v(t) + X'_i(t)\cdot\frac{Q_{iv}}{\tau+\sum_l Q_{il}} + Q_{jv} \\
&\qquad\qquad + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} - \sum_m Q_{vm}\Bigg\} \\
&= UI'_v(t+1).
\end{aligned}
$$

So we have proven $UI'_v(t+1) \leq UI''_v(t+1), \forall v \neq i,j,k$.

For the location i, we have

$$
\begin{aligned}
UI'_i(t+1) &= \max\left\{0,\ UI'_i(t) + \sum_m Y'_{mi}(t) - \sum_m Q_{im}\right\} \\
&= \max\left\{0,\ UI'_i(t) + X'_j(t)\cdot\frac{Q_{ji}}{\sum_l Q_{jl}} + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}} - \sum_m Q_{im}\right\} \\
&= \max\left\{0,\ UI'_i(t) + \left(\sum_l Q_{jl}\right)\cdot\frac{Q_{ji}}{\sum_l Q_{jl}} + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}} - \sum_m Q_{im}\right\} \\
&= \max\left\{0,\ UI'_i(t) + Q_{ji} + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}} - \sum_m Q_{im}\right\}
\end{aligned}
$$

and

$$
UI''_i(t+1) = \max\left\{0,\ UI''_i(t) + \sum_m Y''_{mi}(t) - \sum_m Q_{im}\right\}
$$

$$
\begin{aligned}
&= \max\left\{0,\ UI_i''(t) + X_j''(t)\cdot\frac{Q_{ji}}{\tau+\sum_l Q_{jl}} + \sum_{m\neq i,j} X_m''(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}} - \sum_m Q_{im}\right\}\\
&= \max\left\{0,\ UI_i''(t) + \left(\tau+\sum_l Q_{jl}\right)\cdot\frac{Q_{ji}}{\tau+\sum_l Q_{jl}} + \sum_{m\neq i,j} X_m''(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}} - \sum_m Q_{im}\right\}\\
&= \max\left\{0,\ UI_i''(t) + Q_{ji} + \sum_{m\neq i,j} X_m''(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}} - \sum_m Q_{im}\right\}.
\end{aligned}
$$

Because we have

$$
\begin{aligned}
&UI_i'(t) \le UI_i''(t),\\
&X_v'(t) \le X_v''(t), \forall v \neq i, j,
\end{aligned}
$$

we have

$$
\begin{aligned}
UI_i'(t+1) &= \max\left\{0,\ UI_i'(t) + Q_{ji} + \sum_{m\neq i,j} X_m'(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}} - \sum_m Q_{im}\right\}\\
&\le \max\left\{0,\ UI_i''(t) + Q_{ji} + \sum_{m\neq i,j} X_m''(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}} - \sum_m Q_{im}\right\}\\
&= UI_i''(t+1).
\end{aligned}
$$

So we have proven $UI_i'(t+1) \le UI_i''(t+1)$.

For the location k, we have

$$
\begin{aligned}
UI_k'(t+1) &= \max\left\{0,\ UI_k'(t) + \sum_m Y_{mk}'(t) - \sum_m Q_{km}\right\}\\
&= \max\left\{0,\ UI_k'(t) + X_i'(t)\cdot\frac{\tau+Q_{ik}}{\tau+\sum_l Q_{il}} + X_j'(t)\cdot\frac{Q_{jk}}{\sum_l Q_{jl}}\right.
\end{aligned}
$$

$$
\begin{aligned}
&\qquad + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} - \sum_m Q_{km}\Bigg\} \\
&= \max\Bigg\{0,\ UI'_k(t) + X'_i(t)\cdot\frac{\tau+Q_{ik}}{\tau+\sum_l Q_{il}} + \left(\sum_l Q_{jl}\right)\cdot\frac{Q_{jk}}{\sum_l Q_{jl}} \\
&\qquad + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} - \sum_m Q_{km}\Bigg\} \\
&= \max\Bigg\{0,\ UI'_k(t) + X'_i(t)\cdot\frac{\tau+Q_{ik}}{\tau+\sum_l Q_{il}} + Q_{jk} \\
&\qquad + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} - \sum_m Q_{km}\Bigg\}
\end{aligned}
$$

and

$$
\begin{aligned}
UI''_k(t+1) &= \max\left\{0,\ UI''_k(t) + \sum_m Y''_{mk}(t) - \sum_m Q_{km}\right\} \\
&= \max\Bigg\{0,\ UI''_k(t) + X''_i(t)\cdot\frac{Q_{ik}}{\sum_l Q_{il}} + X''_j(t)\cdot\frac{\tau+Q_{jk}}{\tau+\sum_l Q_{jl}} \\
&\qquad + \sum_{m\neq i,j} X''_m(t)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} - \sum_m Q_{km}\Bigg\} \\
&= \max\Bigg\{0,\ UI''_k(t) + X''_i(t)\cdot\frac{Q_{ik}}{\sum_l Q_{il}} + \left(\tau+\sum_l Q_{jl}\right)\cdot\frac{\tau+Q_{jk}}{\tau+\sum_l Q_{jl}} \\
&\qquad + \sum_{m\neq i,j} X''_m(t)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} - \sum_m Q_{km}\Bigg\} \\
&= \max\Bigg\{0,\ UI''_k(t) + X''_i(t)\cdot\frac{Q_{ik}}{\sum_l Q_{il}} + \tau + Q_{jk} \\
&\qquad + \sum_{m\neq i,j} X''_m(t)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} - \sum_m Q_{km}\Bigg\}
\end{aligned}
$$

Because we have

$$X_i'(t) \le \tau + \sum_l Q_{il} \;\Rightarrow\; X_i'(t) \cdot \frac{\tau}{\tau + \sum_l Q_{il}} - \tau \le 0$$
$$\frac{X_i'(t)}{\tau + \sum_{m=1}^n Q_{im}} \le \frac{X_i''(t)}{\sum_{m=1}^n Q_{im}} \;\Rightarrow\; X_i'(t) \cdot \frac{Q_{ik}}{\tau + \sum_l Q_{il}} - X_i''(t) \cdot \frac{Q_{ik}}{\sum_l Q_{il}} \le 0,$$

we have

$$X_i'(t) \cdot \frac{\tau + Q_{ik}}{\tau + \sum_l Q_{il}} \le X_i''(t) \cdot \frac{Q_{ik}}{\sum_l Q_{il}} + \tau.$$

Together with $X_v'(t) \le X_v''(t),\ \forall v \ne i, j$, we have

$$\begin{aligned} UI_k'(t+1) =& \max \left\{ 0,\ UI_k'(t) + X_i'(t) \cdot \frac{\tau + Q_{ik}}{\tau + \sum_l Q_{il}} + Q_{jk} \right. \\ & \left. + \sum_{m \ne i,j} X_m'(t) \cdot \frac{Q_{mk}}{\sum_l Q_{ml}} - \sum_m Q_{km} \right\} \\ \le& \max \left\{ 0,\ UI_k''(t) + X_i''(t) \cdot \frac{Q_{ik}}{\sum_l Q_{il}} + \tau + Q_{jk} \right. \\ & \left. + \sum_{m \ne i,j} X_m''(t) \cdot \frac{Q_{mk}}{\sum_l Q_{ml}} - \sum_m Q_{km} \right\} \\ =& UI_k''(t+1). \end{aligned}$$

So we have proven $UI_k'(t+1) \le UI_k''(t+1)$.

For location $v \ne i, j, k$, we have

$$X_v'(t+1) = \min \left\{ \sum_m Q_{vm},\ UI_v'(t) + \sum_m Y_{mv}'(t) \right\}$$

$$
\begin{aligned}
&= \min\left\{ \sum_m Q_{vm},\ UI'_v(t) + X'_i(t)\cdot\frac{Q_{iv}}{\tau+\sum_l Q_{il}} + X'_j(t)\cdot\frac{Q_{jv}}{\sum_l Q_{jl}} + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{vm},\ UI'_v(t) + X'_i(t)\cdot\frac{Q_{iv}}{\tau+\sum_l Q_{il}} + \left(\sum_l Q_{jl}\right)\cdot\frac{Q_{jv}}{\sum_l Q_{jl}} \right. \\
&\qquad\qquad \left. + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{vm},\ UI'_v(t) + X'_i(t)\cdot\frac{Q_{iv}}{\tau+\sum_l Q_{il}} + Q_{jv} + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} \right\},
\end{aligned}
$$

and

$$
\begin{aligned}
X''_v(t+1) &= \min\left\{ \sum_m Q_{vm},\ UI''_v(t) + \sum_m Y''_{mv}(t) \right\} \\
&= \min\left\{ \sum_m Q_{vm},\ UI''_v(t) + X''_j(t)\cdot\frac{Q_{jv}}{\tau+\sum_l Q_{jl}} + \sum_{m\neq j} X''_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{vm},\ UI''_v(t) + \left(\tau+\sum_l Q_{jl}\right)\cdot\frac{Q_{jv}}{\tau+\sum_l Q_{jl}} + \sum_{m\neq j} X''_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{vm},\ UI''_v(t) + Q_{jv} + X''_i(t)\cdot\frac{Q_{iv}}{\sum_l Q_{il}} + \sum_{m\neq i,j} X''_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} \right\},
\end{aligned}
$$

Because $\frac{X'_i(t)}{\tau+\sum_{m=1}^n Q_{im}} \leq \frac{X''_i(t)}{\sum_{m=1}^n Q_{im}}$ and $UI'_v(t) \leq UI''_v(t)$, we then have

$$
\begin{aligned}
X'_v(t+1) &= \min\left\{ \sum_m Q_{vm},\ UI'_v(t) + X'_i(t)\cdot\frac{Q_{iv}}{\tau+\sum_l Q_{il}} + Q_{jv} + \sum_{m\neq i,j} X'_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} \right\} \\
&\leq \min\left\{ \sum_m Q_{vm},\ UI''_v(t) + Q_{jv} + X''_i(t)\cdot\frac{Q_{iv}}{\sum_l Q_{il}} + \sum_{m\neq i,j} X''_m(t)\cdot\frac{Q_{mv}}{\sum_l Q_{ml}} \right\} \\
&= X''_v(t+1).
\end{aligned}
$$

So we have proven $X'_v(t+1) \leq X''_v(t+1), \forall v \neq i,j,k$.

For the location i, we have

$$\begin{aligned}
X_i'(t+1) &= \min\left\{\tau+\sum_m Q_{im},\ UI_i'(t)+\sum_m Y_{mi}'(t)\right\}\\
&= \min\left\{\tau+\sum_m Q_{im},\ UI_i'(t)+X_j'(t)\cdot\frac{Q_{ji}}{\sum_l Q_{jl}}+\sum_{m\neq i,j} X_m'(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}}\right\}\\
&= \min\left\{\tau+\sum_m Q_{im},\ UI_i'(t)+\left(\sum_l Q_{jl}\right)\cdot\frac{Q_{ji}}{\sum_l Q_{jl}}+\sum_{m\neq i,j} X_m'(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}}\right\}\\
&= \min\left\{\tau+\sum_m Q_{im},\ UI_i'(t)+Q_{ji}+\sum_{m\neq i,j} X_m'(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}}\right\},
\end{aligned}$$

and

$$\begin{aligned}
X_i''(t+1) &= \min\left\{\sum_m Q_{im},\ UI_i''(t)+\sum_m Y_{mi}''(t)\right\}\\
&= \min\left\{\sum_m Q_{im},\ UI_i''(t)+X_j''(t)\cdot\frac{Q_{ji}}{\tau+\sum_l Q_{jl}}+\sum_{m\neq i,j} X_m''(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}}\right\}\\
&= \min\left\{\sum_m Q_{im},\ UI_i''(t)+\left(\tau+\sum_l Q_{jl}\right)\cdot\frac{Q_{ji}}{\tau+\sum_l Q_{jl}}+\sum_{m\neq i,j} X_m''(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}}\right\}\\
&= \min\left\{\sum_m Q_{im},\ UI_i''(t)+Q_{ji}+\sum_{m\neq i,j} X_m''(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}}\right\}.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\frac{X_i'(t+1)}{\tau+\sum_{m=1}^n Q_{im}}\\
=&\frac{\min\left\{\tau+\sum_m Q_{im},\ UI_i'(t)+Q_{ji}+\sum_{m\neq i,j} X_m'(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}}\right\}}{\tau+\sum_{m=1}^n Q_{im}}\\
=&\min\left\{\frac{\tau+\sum_m Q_{im}}{\tau+\sum_{m=1}^n Q_{im}},\ \frac{UI_i'(t)+Q_{ji}+\sum_{m\neq i,j} X_m'(t)\cdot\frac{Q_{mi}}{\sum_l Q_{ml}}}{\tau+\sum_{m=1}^n Q_{im}}\right\}
\end{aligned}$$

$$
\begin{aligned}
&\leq \min\left\{ \frac{\sum_m Q_{im}}{\sum_{m=1}^n Q_{im}},\ \frac{UI_i''(t)+Q_{ji}+\sum_{m\neq i,j} X_m''(t)\cdot \frac{Q_{mi}}{\sum_l Q_{ml}}}{\sum_{m=1}^n Q_{im}} \right\} \\
&= \frac{\min\left\{ \sum_m Q_{im},\ UI_i''(t)+Q_{ji}+\sum_{m\neq i,j} X_m''(t)\cdot \frac{Q_{mi}}{\sum_l Q_{ml}} \right\}}{\sum_{m=1}^n Q_{im}} \\
&= \frac{X_i''(t+1)}{\sum_{m=1}^n Q_{im}}.
\end{aligned}
$$

So we have proven $\frac{X_i'(t+1)}{\tau+\sum_{m=1}^n Q_{im}} \leq \frac{X_i''(t+1)}{\sum_{m=1}^n Q_{im}}$.

For the location k, we have

$$
\begin{aligned}
X_k'(t+1) &= \min\left\{ \sum_m Q_{km},\ UI_k'(t)+\sum_m Y_{mk}'(t) \right\} \\
&= \min\left\{ \sum_m Q_{km},\ UI_k'(t)+X_i'(t)\cdot\frac{\tau+Q_{ik}}{\tau+\sum_l Q_{il}}+X_j'(t)\cdot\frac{Q_{jk}}{\sum_l Q_{jl}}+\sum_{m\neq i,j} X_m'(t)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{km},\ UI_k'(t)+X_i'(t)\cdot\frac{\tau+Q_{ik}}{\tau+\sum_l Q_{il}}+\left(\sum_l Q_{jl}\right)\cdot\frac{Q_{jk}}{\sum_l Q_{jl}} \right. \\
&\qquad\qquad \left. +\sum_{m\neq i,j} X_m'(t)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{km},\ UI_k'(t)+X_i'(t)\cdot\frac{\tau+Q_{ik}}{\tau+\sum_l Q_{il}}+Q_{jk}+\sum_{m\neq i,j} X_m'(t)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} \right\},
\end{aligned}
$$

and

$$
\begin{aligned}
X_k''(t+1) &= \min\left\{ \sum_m Q_{km},\ UI_k''(t)+\sum_m Y_{mk}''(t) \right\} \\
&= \min\left\{ \sum_m Q_{km},\ UI_k''(t)+X_j''(t)\cdot\frac{\tau+Q_{jk}}{\tau+\sum_l Q_{jl}}+\sum_{m\neq j} X_m''(t)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} \right\} \\
&= \min\left\{ \sum_m Q_{km},\ UI_k''(t)+\left(\tau+\sum_l Q_{jl}\right)\cdot\frac{\tau+Q_{jk}}{\tau+\sum_l Q_{jl}}+\sum_{m\neq j} X_m''(t)\cdot\frac{Q_{mk}}{\sum_l Q_{ml}} \right\}
\end{aligned}
$$

$$= \min \left\{ \sum_m Q_{km},\ UI''_k(t) + \tau + Q_{jk} + X''_i(t) \cdot \frac{Q_{ik}}{\sum_l Q_{il}} + \sum_{m \neq i,j} X''_m(t) \cdot \frac{Q_{mk}}{\sum_l Q_{ml}} \right\},$$

Because we have

$$X'_i(t) \leq \tau + \sum_l Q_{il} \ \Rightarrow\ X'_i(t) \cdot \frac{\tau}{\tau + \sum_l Q_{il}} - \tau \leq 0$$
$$\frac{X'_i(t)}{\tau + \sum_{m=1}^n Q_{im}} \leq \frac{X''_i(t)}{\sum_{m=1}^n Q_{im}} \ \Rightarrow\ X'_i(t) \cdot \frac{Q_{ik}}{\tau + \sum_l Q_{il}} - X''_i(t) \cdot \frac{Q_{ik}}{\sum_l Q_{il}} \leq 0,$$

we have

$$X'_i(t) \cdot \frac{\tau + Q_{ik}}{\tau + \sum_l Q_{il}} \leq X''_i(t) \cdot \frac{Q_{ik}}{\sum_l Q_{il}} + \tau.$$

Together with $UI'_k(t) \leq UI''_k(t)$ and $X'_v(t) \leq X''_v(t),\ \forall v \neq i, j$, we have

$$\begin{aligned}
& X'_k(t+1) \\
=& \min \left\{ \sum_m Q_{km},\ UI'_k(t) + X'_i(t) \cdot \frac{\tau + Q_{ik}}{\tau + \sum_l Q_{il}} + Q_{jk} + \sum_{m \neq i,j} X'_m(t) \cdot \frac{Q_{mk}}{\sum_l Q_{ml}} \right\} \\
\leq& \min \left\{ \sum_m Q_{km},\ UI''_k(t) + \tau + Q_{jk} + X''_i(t) \cdot \frac{Q_{ik}}{\sum_l Q_{il}} + \sum_{m \neq i,j} X''_m(t) \cdot \frac{Q_{mk}}{\sum_l Q_{ml}} \right\} \\
=& X''_k(t+1).
\end{aligned}$$

So we have proven that the relation (A.4b), (A.4c) and (A.4d) hold for the time period $t+1$. Therefore, we have proven (A.4) hold.

Part 2. We have

$$\widehat{x}_i^* = \lim_{t \to \infty} X''_i(t) \geq \lim_{t \to \infty} X'_i(t) \cdot \frac{\sum_{m=1}^n Q_{im}}{\tau + \sum_{m=1}^n Q_{im}} = \widetilde{x}_i^* \cdot \frac{\sum_{m=1}^n Q_{im}}{\tau + \sum_{m=1}^n Q_{im}},$$

$$\widehat{x}_v^* = \lim_{t\to\infty} X_v''(t) \geq \lim_{t\to\infty} X_v'(t) = \widetilde{x}_v^*,\ \forall v \neq i, j,$$

Together with $\widehat{x}_j^* = \tau + \sum_{m=1}^n Q_{jm}$, $\widetilde{x}_j^* = \sum_{m=1}^n Q_{jm}$, we have

$$\begin{aligned}
TF'' =& \text{Total network flow in Scenario 3} \\
=& \widehat{x}_i^* + \widehat{x}_j^* + \sum_{v\neq i,j} \widehat{x}_v^* \\
\geq& \widetilde{x}_i^* \cdot \frac{\sum_{m=1}^n Q_{im}}{\tau + \sum_{m=1}^n Q_{im}} + \tau + \widetilde{x}_j^* + \sum_{v\neq i,j} \widetilde{x}_v^* \\
=& \widetilde{x}_i^* \cdot \left(\frac{\sum_{m=1}^n Q_{im}}{\tau + \sum_{m=1}^n Q_{im}} - 1 \right) + \tau + \widetilde{x}_i^* + \widetilde{x}_j^* + \sum_{v\neq i,j} \widetilde{x}_v^* \\
=& -\widetilde{x}_i^* \cdot \left(\frac{\tau}{\tau + \sum_{m=1}^n Q_{im}} \right) + \tau + \widetilde{x}_i^* + \widetilde{x}_j^* + \sum_{v\neq i,j} \widetilde{x}_v^* \\
=& \left[1 - \frac{\widetilde{x}_i^*}{\tau + \sum_{m=1}^n Q_{im}} \right] \cdot \tau + \widetilde{x}_i^* + \widetilde{x}_j^* + \sum_{v\neq i,j} \widetilde{x}_v^* \\
\geq& \left[1 - \frac{\widehat{x}_i^*}{\sum_{m=1}^n Q_{im}} \right] \cdot \tau + \widetilde{x}_i^* + \widetilde{x}_j^* + \sum_{v\neq i,j} \widetilde{x}_v^*.
\end{aligned}$$

From the proof of Proposition 6, we know that

$$\widehat{x}_j^* = \tau + \sum_{m=1}^n Q_{jm},$$
$$\widehat{x}_v^* \geq x_v^*,\ \forall v \neq j$$

From the proof of Proposition 3, we know that

$$\sum_v (\widehat{x}_v^* - x_v^*) \leq \tau \cdot \left(1 + \frac{1}{\min_k \alpha_{kj}} \right)$$

We then have

$$\begin{aligned}
&\widehat{x}_i^* \\
\leq & x_i^* - \sum_{v \neq i} (\widehat{x}_v^* - x_v^*) + \tau \cdot \left(1 + \frac{1}{\min_k \alpha_{kj}}\right) \\
\leq & x_i^* + \tau \cdot \left(1 + \frac{1}{\min_k \alpha_{kj}}\right)
\end{aligned}$$

Then we have

$$\begin{aligned}
TF'' = & \text{Total network flow in Scenario 3} \\
\geq & \left[1 - \frac{\widehat{x}_i^*}{\sum_{m=1}^n Q_{im}}\right] \cdot \tau + \widetilde{x}_i^* + \widetilde{x}_j^* + \sum_{v \neq i,j} \widetilde{x}_v^* \\
\geq & \left\{1 - \frac{1}{\sum_{m=1}^n Q_{im}} \cdot \left[x_i^* + \tau \cdot \left(1 + \frac{1}{\min_k \alpha_{kj}}\right)\right]\right\} \cdot \tau + \widetilde{x}_i^* + \widetilde{x}_j^* + \sum_{v \neq i,j} \widetilde{x}_v^* \\
= & \tau \cdot \left(1 - \frac{x_i^*}{\sum_{m=1}^n Q_{im}}\right) - \tau^2 \cdot \frac{1}{\sum_{m=1}^n Q_{im}} \cdot \left(1 + \frac{1}{\min_k \alpha_{kj}}\right) + \sum_v \widetilde{x}_v^* \\
= & \tau \cdot \left(1 - \frac{x_i^*}{\sum_{m=1}^n Q_{im}}\right) - \tau^2 \cdot \frac{1}{\sum_{m=1}^n Q_{im}} \cdot \left(1 + \frac{1}{\min_k \alpha_{kj}}\right) + TF'
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\lambda_{j \to k} - \lambda_{i \to k} \\
= & \lim_{\tau \to 0} \frac{TF'' - TF}{\tau} - \lim_{\tau \to 0} \frac{TF' - TF}{\tau} \\
= & \lim_{\tau \to 0} \frac{TF'' - TF'}{\tau} \\
\geq & \lim_{\tau \to 0} \frac{\tau \cdot \left(1 - \frac{x_i^*}{\sum_{m=1}^n Q_{im}}\right) - \tau^2 \cdot \frac{1}{\sum_{m=1}^n Q_{im}} \cdot \left(1 + \frac{1}{\min_k \alpha_{kj}}\right)}{\tau}
\end{aligned}$$

$$
\begin{aligned}
&\geq 1 - \frac{x_i^*}{\sum_{m=1}^n Q_{im}} - \lim_{\tau \to 0} \tau \cdot \frac{1}{\sum_{m=1}^n Q_{im}} \cdot \left(1 + \frac{1}{\min_k \alpha_{kj}}\right) \\
&= 1 - \frac{x_i^*}{\sum_{m=1}^n Q_{im}} > 0
\end{aligned}
$$

Hence proven. □

A.5 Proof of Proposition 5.

First we prove the following lemma.

Lemma 2 (System with Saturated Initialization)**.** *Assume a complete network with the demand pattern $\boldsymbol{Q}$, and denote the equilibrium flow as $\boldsymbol{x}^*$. If we place no less than $\sum_{k=1}^n Q_{jk}$ vehicles at each location j at the beginning of the time $t = 1$, then $X_j(t) \geq x_j^*, \forall j = 1, \cdots, n, \forall t \geq 1$.*

Proof. Proof of Lemma 2 We prove the proposition by induction.

For $t = 1$, we have $UI_j(1) = 0$, $X_j(t) \geq \sum_{k=1}^n Q_{jk} \geq x_j^*$, $\forall j = 1, \cdots, n$. At t=2, we have

$$
\begin{aligned}
X_j(2) &= \min\left\{ \sum_{k=1}^n Q_{jk},\ UI_j(1) + \sum_{i=1}^n X_i(1) \cdot \frac{Q_{ij}}{\sum_{k=1}^n Q_{ik}} \right\} \\
&= \min\left\{ \sum_{k=1}^n Q_{jk},\ \sum_{i=1}^n X_i(1) \cdot \frac{Q_{ij}}{\sum_{k=1}^n Q_{ik}} \right\} \\
&\geq \min\left\{ \sum_{k=1}^n Q_{jk},\ \sum_{i=1}^n x_i^* \cdot \frac{Q_{ij}}{\sum_{k=1}^n Q_{ik}} \right\} \\
&\geq x_j^*
\end{aligned}
$$

The last inequality holds because both $\sum_{k=1}^n Q_{jk} \geq x_j^*$ and $\sum_{i=1}^n x_i^* \cdot \frac{Q_{ij}}{\sum_{k=1}^n Q_{ik}} = x_j^*$ hold. Therefore, the result holds for both $t = 1$ and $t = 2$. Now suppose $X_j(t) \geq x_j^*, \forall j = 1, \cdots, n$ holds for the time

period t. Next we prove that the result also holds for the time period $t+1$. We have

$$\begin{aligned}
X_j(t+1) &= \min\left\{\sum_{k=1}^{n} Q_{jk},\ UI_j(t) + \sum_{i=1}^{n} X_i(t)\cdot\frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}}\right\} \\
&\geq \min\left\{\sum_{k=1}^{n} Q_{jk},\ \sum_{i=1}^{n} X_i(t)\cdot\frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}}\right\} \\
&\geq \min\left\{\sum_{k=1}^{n} Q_{jk},\ \sum_{i=1}^{n} x_i^*\cdot\frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}}\right\} \\
&\geq x_j^*
\end{aligned}$$

Therefore, $X_j(t+1) \geq x_j^*, \forall j = 1,\cdots,n$ holds. Hence proven. □

Then we prove Proposition 5.

Proof. Proof of Proposition 5 (1) We just need to prove $\widetilde{x}_j^* \leq x_j^*, \forall j = 1,\cdots,n$, which leads to the result $R(\boldsymbol{Q}+\Delta) \leq R(\boldsymbol{Q})$ because the optimal prices for both demand pattern $\boldsymbol{Q}$ and $\boldsymbol{Q}+\Delta$ are the same.

For a fixed network, we consider the following two cases.

1. Case 1: the demand pattern is $\boldsymbol{Q}$. At the beginning of $t=1$, we place $\sum_{k=1}^{n} Q_{jk}$ vehicles at each location j. Denote the flow from location i to location j during the time period t as $Y_{ij}(t)$. Denote the outbound flow from location j during time period t as $X_j(t)$. We use $UI_j(t)$ to denote the number of vehicles that are not used by any customer during the period t and stay at location j during the period t.

2. Case 2: the demand pattern is $\boldsymbol{Q}+\Delta$. At the beginning of $t=1$, we place $\sum_{k=1}^{n}(Q_{jk}+\delta_{jk})$ vehicles at each location j. Denote the flow from location i to location j during the time period t as $Y'_{ij}(t)$. Denote the outbound flow from location j during time

period t as $X'_j(t)$. We use $UI'_j(t)$ to denote the number of vehicles that are not used by any customer during the period t and stay at location j during the period t.

According to Proposition 2, we know the system state $X_j(t)$s will evolve into the corresponding equilibrium flow x^*_js and the system state $X'_j(t)$s will evolve into the corresponding equilibrium flow $\widetilde{x}^*_j$s. According to Lemma 2, we know $X_j(t) \geq x^*_j, \forall j \in J_S(\boldsymbol{Q})$. We also know that if $j \in J_S(\boldsymbol{Q})$, then $X_j(t) \leq \sum_{k=1}^n Q_{jk} = x^*_j$. Therefore, we have

$$X_j(t) = \sum_{k=1}^{n} Q_{jk} = x^*_j, \forall t \geq 1, \forall j \in J_S(\boldsymbol{Q}). \tag{A.5a}$$

For all $i, j = 1, \cdots, n, t \geq 1$, we have the following relations.

$$Y_{ij}(t) = X_i(t) \cdot \frac{Q_{ij}}{\sum_{k=1}^n Q_{ik}} \tag{A.6a}$$

$$Y'_{ij}(t) = X'_i(t) \cdot \frac{Q_{ij} + \delta_{ij}}{\sum_{k=1}^n (Q_{ik} + \delta_{ik})} \tag{A.6b}$$

$$X_j(t+1) = \min \left\{ \sum_{k=1}^{n} Q_{jk},\ UI_j(t) + \sum_i Y_{ij}(t) \right\} \tag{A.6c}$$

$$X'_j(t+1) = \min \left\{ \sum_{k=1}^{n} (Q_{jk} + \delta_{jk}),\ UI'_j(t) + \sum_i Y'_{ij}(t) \right\} \tag{A.6d}$$

$$UI_j(t+1) = \max \left\{ 0,\ UI_j(t) + \sum_i Y_{ij}(t) - \sum_{k=1}^{n} Q_{jk} \right\} \tag{A.6e}$$

$$UI'_j(t+1) = \max \left\{ 0,\ UI'_j(t) + \sum_i Y'_{ij}(t) - \sum_{k=1}^{n} (Q_{jk} + \delta_{jk}) \right\} \tag{A.6f}$$

We know $\delta_{jk} = 0, \forall j \in S_1, k = 1, \cdots, n$ because otherwise the set A_1 is not empty. Next we prove the following results by induction.

$$X'_j(t) \leq X_j(t), \forall j \in S_1, \forall t \geq 2 \tag{A.7a}$$

$$UI'_j(t) \leq UI_j(t), \forall j \in \{1,\cdots,n\} \setminus S_1, \forall t \geq 2 \tag{A.7b}$$

$$Y'_{ij}(t) \leq Y_{ij}(t), \forall i = 1,\cdots,n, j \in \{1,\cdots,n\} \setminus S_1, \forall t \geq 2 \tag{A.7c}$$

We know that

$$X_j(1) = \sum_k Q_{jk},\ X'_j(1) = \sum_k (Q_{jk} + \delta_{jk}), \forall j = 1,\cdots,n$$

$$X_j(1) = X'_j(1) = \sum_k Q_{jk}, \forall j \in S_1$$

$$UI_j(1) = 0,\ UI'_j(1) = 0, \forall j = 1,\cdots,n$$

We also have

$$Y'_{ij}(1) = Q_{ij} = Y_{ij}(1), \forall i = 1,\cdots,n, j \in \{1,\cdots,n\} \setminus S_1 \tag{A.8a}$$

We then prove the relation (A.7) for $t = 2$. If $j \in S_1$, then $\delta_{jk} = 0, \forall k = 1,\cdots,n$. So we have

$$\begin{aligned}
X'_j(2) =& \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \delta_{jk}),\ UI'_j(1) + \sum_{i=1}^{n} X'_i(1) \cdot \frac{Q_{ij} + \delta_{ij}}{\sum_{k=1}^{n} (Q_{ik} + \delta_{ik})} \right\} \\
=& \min\left\{ \sum_{k=1}^{n} Q_{jk},\ \sum_{i=1}^{n} X'_i(1) \cdot \frac{Q_{ij} + \delta_{ij}}{\sum_{k=1}^{n} (Q_{ik} + \delta_{ik})} \right\} \\
\leq& \sum_{k=1}^{n} Q_{jk} \\
=& X_j(2)
\end{aligned}$$

The last equality is due to the relation (A.5). If $j \in \{1,\cdots,n\} \setminus S_1$, then $\delta_{ij} = 0, \delta_{ji} \geq 0, \forall i =$

$1, \cdots, n$. We have

$$\begin{aligned}
UI'_j(2) =& \max\left\{0,\ UI'_j(1) + \sum_i Y'_{ij}(1) - \sum_{k=1}^{n}(Q_{jk} + \delta_{jk})\right\} \\
\leq& \max\left\{0,\ UI'_j(1) + \sum_i Y'_{ij}(1) - \sum_{k=1}^{n} Q_{jk}\right\} \\
\leq& \max\left\{0,\ UI_j(1) + \sum_i Y_{ij}(1) - \sum_{k=1}^{n} Q_{jk}\right\} \\
=& UI_j(2),
\end{aligned}$$

where in the last inequality we used the relation (A.8).

If $j \in \{1, \cdots, n\} \setminus S_1$, then $\delta_{ij} = 0, \delta_{ji} \geq 0, \forall i = 1, \cdots, n$. We prove $Y'_{ij}(2) \leq Y_{ij}(2), \forall i = 1, \cdots, n$ by discussing two cases: the first case $i \in S_1$ and the second case $i \in \{1, \cdots, n\} \setminus S_1$. For the case $i \in S_1$, we know $X_i(2) = \sum_{k=1}^{n} Q_{ik} = x_i^*$ because of the relation (A.5). So we have $Y_{ij}(2) = X_i(2) \cdot \frac{Q_{ij}}{\sum_k Q_{ik}} = Q_{ij}$. Then we have $Y'_{ij}(2) \leq Q_{ij} = Y_{ij}(2)$. For the case $i \in \{1, \cdots, n\} \setminus S_1$, we have

$$\begin{aligned}
Y'_{ij}(2) =& X'_i(2) \cdot \frac{Q_{ij} + \delta_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \\
=& \frac{Q_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \cdot \min\left\{\sum_{m=1}^{n}(Q_{im} + \delta_{im}),\ UI'_i(1) + \sum_l Y'_{li}(1)\right\} \\
\leq& \frac{Q_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \cdot \min\left\{\sum_{m=1}^{n}(Q_{im} + \delta_{im}),\ UI_i(1) + \sum_l Y_{li}(1)\right\} \\
=& \min\left\{\frac{Q_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \cdot \sum_{m=1}^{n}(Q_{im} + \delta_{im}),\ \frac{Q_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \cdot \left(UI_i(1) + \sum_l Y_{li}(1)\right)\right\} \\
=& \min\left\{\frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}} \cdot \sum_{m=1}^{n} Q_{im},\ \frac{Q_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \cdot \left(UI_i(1) + \sum_l Y_{li}(1)\right)\right\} \\
\leq& \min\left\{\frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}} \cdot \sum_{m=1}^{n} Q_{im},\ \frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}} \cdot \left(UI_i(1) + \sum_l Y_{li}(1)\right)\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}} \cdot \min\left\{ \sum_{m=1}^{n} Q_{im},\ UI_i(1) + \sum_{l} Y_{li}(1) \right\} \\
&= \frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}} \cdot X_i(2) \\
&= Y_{ij}(2)
\end{aligned}
$$

Therefore, we have proven $Y'_{ij}(2) \leq Y_{ij}(2), \forall i = 1, \cdots, n$.

Now assume the relation (A.7) holds for time period $t \geq 2$. We next prove it also holds for time period $t+1$. If $j \in S_1$, then $\delta_{jk} = 0, \forall k = 1, \cdots, n$. So we have

$$
\begin{aligned}
X'_j(t+1) &= \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \delta_{jk}),\ UI'_j(t) + \sum_{i=1}^{n} X'_i(t) \cdot \frac{Q_{ij} + \delta_{ij}}{\sum_{k=1}^{n} (Q_{ik} + \delta_{ik})} \right\} \\
&= \min\left\{ \sum_{k=1}^{n} Q_{jk},\ UI'_j(t) + \sum_{i=1}^{n} X'_i(t) \cdot \frac{Q_{ij} + \delta_{ij}}{\sum_{k=1}^{n} (Q_{ik} + \delta_{ik})} \right\} \\
&\leq \sum_{k=1}^{n} Q_{jk} \\
&= X_j(t+1)
\end{aligned}
$$

The last equality is due to the relation (A.5). If $j \in \{1, \cdots, n\} \setminus S_1$, then $\delta_{ij} = 0, \delta_{ji} \geq 0, \forall i = 1, \cdots, n$. We have

$$
\begin{aligned}
UI'_j(t+1) &= \max\left\{ 0,\ UI'_j(t) + \sum_{i} Y'_{ij}(t) - \sum_{k=1}^{n} (Q_{jk} + \delta_{jk}) \right\} \\
&\leq \max\left\{ 0,\ UI'_j(t) + \sum_{i} Y'_{ij}(t) - \sum_{k=1}^{n} Q_{jk} \right\} \\
&\leq \max\left\{ 0,\ UI_j(t) + \sum_{i} Y_{ij}(t) - \sum_{k=1}^{n} Q_{jk} \right\} \\
&= UI_j(t+1)
\end{aligned}
$$

If $j \in \{1,\cdots,n\} \setminus S_1$, then $\delta_{ij} = 0, \delta_{ji} \geq 0, \forall i = 1,\cdots,n$. We prove $Y'_{ij}(t+1) \leq Y_{ij}(t+1), \forall i = 1,\cdots,n$ by discussing two cases: $i \in S_1$ and $i \in \{1,\cdots,n\} \setminus S_1$. For any $i \in S_1$, we know $X_i(t+1) = \sum_{k=1}^{n} Q_{ik} = x_i^*$ because of the relation (A.5). So we have $Y_{ij}(t+1) = X_i(t+1) \cdot \frac{Q_{ij}}{\sum_k Q_{ik}} = Q_{ij}$. Then we have $Y'_{ij}(t+1) \leq Q_{ij} = Y_{ij}(t+1)$. For any $i \in \{1,\cdots,n\} \setminus S_1$, we have

$$
\begin{aligned}
Y'_{ij}(t+1) =& X'_i(t+1) \cdot \frac{Q_{ij} + \delta_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \\
=& \frac{Q_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \cdot \min\left\{ \sum_{m=1}^{n}(Q_{im} + \delta_{im}),\ UI'_i(t) + \sum_l Y'_{li}(t) \right\} \\
\leq& \frac{Q_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \cdot \min\left\{ \sum_{m=1}^{n}(Q_{im} + \delta_{im}),\ UI_i(t) + \sum_l Y_{li}(t) \right\} \\
=& \min\left\{ \frac{Q_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \cdot \sum_{m=1}^{n}(Q_{im} + \delta_{im}),\ \frac{Q_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \cdot \left(UI_i(t) + \sum_l Y_{li}(t) \right) \right\} \\
=& \min\left\{ \frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}} \cdot \sum_{m=1}^{n} Q_{im},\ \frac{Q_{ij}}{\sum_{k=1}^{n}(Q_{ik} + \delta_{ik})} \cdot \left(UI_i(t) + \sum_l Y_{li}(t) \right) \right\} \\
\leq& \min\left\{ \frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}} \cdot \sum_{m=1}^{n} Q_{im},\ \frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}} \cdot \left(UI_i(t) + \sum_l Y_{li}(t) \right) \right\} \\
=& \frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}} \cdot \min\left\{ \sum_{m=1}^{n} Q_{im},\ UI_i(t) + \sum_l Y_{li}(t) \right\} \\
=& \frac{Q_{ij}}{\sum_{k=1}^{n} Q_{ik}} \cdot X_i(t+1) \\
=& Y_{ij}(t+1)
\end{aligned}
$$

We have proven that the relation (A.7) holds for time period $t+1$. Therefore, the relation (A.7) is proven.

As $t \to \infty$, we have

$$X_j(t) \to x_j^*,\ X'_j(t) \to \widetilde{x}_j^*,\ \forall j \in S_1,$$

$$\sum_i Y_{ij}(t) \to x_j^*,\ \sum_i Y'_{ij}(t) \to \widetilde{x}_j^*,\ j \in \{1,\cdots,n\} \setminus S_1.$$

Because we have the relation (A.7), we have $\widetilde{x}_j^* \leq x_j^*, \forall j = 1,\cdots,n$.

(2) We just need to prove $\widetilde{x}_j^* < x_j^*, \forall j \in \{1,\cdots,n\} \setminus S_1$. Assume a node $j_1 \in A_2$. There must be j_2 and j_3 such that $\delta_{j_2,j_3} > 0$. Obviously j_1 cannot be j_2 or j_3 because otherwise $j_1 \notin A_2$. Because $j_1 \in A_2$, we have $\delta_{j,j_1} = 0, \forall j = 1,\cdots,n$.

$$\begin{aligned}
\widetilde{x}_{j_1}^* =& \widetilde{x}_{j_2}^* \cdot \frac{Q_{j_2,j_1} + \delta_{j_2,j_1}}{\sum_{k=1}^n (Q_{j_2,k} + \delta_{j_2,k})} + \sum_{i \neq j_2, i=1}^n \widetilde{x}_i^* \cdot \frac{Q_{i,j_1} + \delta_{i,j_1}}{\sum_{k=1}^n (Q_{ik} + \delta_{i,k})} \\
=& \widetilde{x}_{j_2}^* \cdot \frac{Q_{j_2,j_1}}{\sum_{k=1}^n (Q_{j_2,k} + \delta_{j_2,k})} + \sum_{i \neq j_2, i=1}^n \widetilde{x}_i^* \cdot \frac{Q_{i,j_1}}{\sum_{k=1}^n (Q_{ik} + \delta_{i,k})} \\
\leq& \widetilde{x}_{j_2}^* \cdot \frac{Q_{j_2,j_1}}{\delta_{j_2,j_3} + \sum_{k=1}^n Q_{j_2,k}} + \sum_{i \neq j_2, i=1}^n \widetilde{x}_i^* \cdot \frac{Q_{i,j_1}}{\sum_{k=1}^n Q_{ik}} \\
<& \widetilde{x}_{j_2}^* \cdot \frac{Q_{j_2,j_1}}{\sum_{k=1}^n Q_{j_2,k}} + \sum_{i \neq j_2, i=1}^n \widetilde{x}_i^* \cdot \frac{Q_{i,j_1}}{\sum_{k=1}^n Q_{ik}} \\
\leq& x_{j_2}^* \cdot \frac{Q_{j_2,j_1}}{\sum_{k=1}^n Q_{j_2,k}} + \sum_{i \neq j_2, i=1}^n x_i^* \cdot \frac{Q_{i,j_1}}{\sum_{k=1}^n Q_{ik}} \\
=& x_{j_1}^*,
\end{aligned}$$

where the last inequality is due to the result (1). For any node $j \in \{1,\cdots,n\} \setminus S_1$, if $j \neq j_1$, we have $\delta_{i,j} = 0, \forall i = 1,\cdots,n$. Then

$$\begin{aligned}
\widetilde{x}_j^* =& \widetilde{x}_{j_1}^* \cdot \frac{Q_{j_1,j} + \delta_{j_1,j}}{\sum_{k=1}^n (Q_{j_1,k} + \delta_{j_1,k})} + \sum_{i \neq j_1, i=1}^n \widetilde{x}_i^* \cdot \frac{Q_{i,j} + \delta_{i,j}}{\sum_{k=1}^n (Q_{ik} + \delta_{i,k})} \\
=& \widetilde{x}_{j_1}^* \cdot \frac{Q_{j_1,j}}{\sum_{k=1}^n (Q_{j_1,k} + \delta_{j_1,k})} + \sum_{i \neq j_1, i=1}^n \widetilde{x}_i^* \cdot \frac{Q_{i,j}}{\sum_{k=1}^n (Q_{ik} + \delta_{i,k})} \\
\leq& \widetilde{x}_{j_1}^* \cdot \frac{Q_{j_1,j}}{\sum_{k=1}^n Q_{j_1,k}} + \sum_{i \neq j_1, i=1}^n \widetilde{x}_i^* \cdot \frac{Q_{i,j}}{\sum_{k=1}^n Q_{ik}}
\end{aligned}$$

$$
\begin{aligned}
&< x^*_{j_1} \cdot \frac{Q_{j_1,j}}{\sum_{k=1}^{n} Q_{j_1,k}} + \sum_{i \neq j_1, i=1}^{n} x^*_i \cdot \frac{Q_{i,j}}{\sum_{k=1}^{n} Q_{ik}} \\
&= x^*_j,
\end{aligned}
$$

Therefore, we have proven $\widetilde{x}^*_j < x^*_j, \forall j \in \{1, \cdots, n\} \setminus S_1$. Hence proven. □

A.6 Proof of Proposition 6.

It is easy to prove the proposition if there are only two locations in the network. In the following, we prove the proposition for networks with at least 3 locations.

Suppose the location j is the only critical location and τ satisfies $0 < \tau \leq \min_{v:\ v \neq j} \left(\sum_{k=1}^{n} Q_{vk} - x^*_v \right) \cdot \alpha_{vj} = \min_{v:\ v \neq j} \left(\sum_{k=1}^{n} Q_{vk} - x^*_v \right) \cdot \alpha'_{vj}$. We consider the following two cases.

1. Scenario 1: at the beginning of each period, there are Q_{iv} potential customers seeking to go from location i to location v. At the beginning of $t = 1$, we place $\sum_{k=1}^{n} Q_{vk}$ vehicles at each location $v \in \{1, \cdots, n\}$. Denote the flow from location i to location v during the time period t as $Y_{iv}(t)$. Denote the outbound flow from location v during time period t as $X_v(t)$. We use $UI_v(t)$ to denote the number of vehicles that are not used by any customer during the period t and stay at location v during the period t.

2. Scenario 2: at the beginning of each period, there are Q_{vk} potential customers seeking to go from location v ($v \neq j$) to location k, and $Q_{jk} + \tau \cdot \varepsilon_{jk}$ potential customers seeking to go from location j to location k. At the beginning of $t = 1$, we place $\sum_{k=1}^{n} Q_{mk}$ vehicles at each location $m \in \{1, \cdots, n\} \setminus \{j\}$, and $\sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk})$ vehicles at the location j. Denote the flow from location i to location v during the time period t as $Y'_{iv}(t)$. Denote the outbound flow from location v during time period t

as $X'_v(t)$. We use $UI'_v(t)$ to denote the number of vehicles that are not used by any customer during the period t and stay at location v during the period t.

For Scenario 1, for each arc (i,k), denote the associated fraction $\frac{Q_{ik}}{\sum_{l=1}^{n} Q_{il}}$ as α_{ik}. For Scenario 2, for each arc (i,k) such that $i \neq j$, denote the associated fraction $\frac{Q_{ik}}{\sum_{l=1}^{n} Q_{il}}$ as α'_{ik}; for Scenario 2, for each arc (j,k), denote the associated fraction $\frac{Q_{jk}+\tau\cdot\varepsilon_{jk}}{\sum_{l=1}^{n}(Q_{jl}+\tau\cdot\varepsilon_{jk})}$ as α'_{jk}.

According to Proposition 2, we know the system state in each scenario will evolve into the corresponding equilibrium flow. Denote the equilibrium flow in Scenario 1 and Scenario 2 as $\boldsymbol{x}^*$ and $\widetilde{\boldsymbol{x}}^*$. According to Lemma 2, we know $X_k(t) \geq x_k^*$ and $X'_k(t) \geq \widetilde{x}_k^*,\ \forall k = 1,\cdots,n,\ \forall t \geq 1$.

We need to prove if $0 < \tau \leq \min_{v:\ v\neq j} \alpha'_{vj} \cdot \left(\sum_{k=1}^{n} Q_{vk} - x_v^*\right)$, then $\widetilde{x}_j^* = \tau + \sum_{k=1}^{n} Q_{jk}$ and $\widetilde{x}_v^* < \sum_{k=1}^{n} Q_{vk}, \forall v \neq j$ hold.

We first prove the following relation holds.

$$X'_j(t) \geq \sum_{k=1}^{n} (Q_{jk} + \tau\cdot\varepsilon_{jk}),\ \forall t \geq 1 \tag{A.9a}$$

$$X'_v(t) > x_v^*,\ \forall v \neq j,\ \forall t \geq 1 \tag{A.9b}$$

In the following, we prove the relation (A.9) by induction. For $t = 1$, we have

$$X'_j(1) = \sum_{k=1}^{n} (Q_{jk} + \tau\cdot\varepsilon_{jk})$$
$$X'_v(1) = \sum_{k=1}^{n} Q_{vk} > x_v^*,\ \forall v \neq j$$

So the relation (A.9) holds for $t = 1$. We next prove it also holds for $t = 2$. We first prove it holds for $t = 2$ for the location j; assume another location in the network is p_0 ($p_0 \neq j$).

$$UI'_j(1) + \sum_i Y'_{ij}(1)$$

$$
\begin{aligned}
&= \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ \sum_{i \neq j} X'_i(1) \cdot \alpha'_{ij} \right\} \\
&= \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ \sum_{i \neq j} \sum_{k \neq i} Q_{ik} \cdot \alpha'_{ij} \right\} \\
&= \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ \sum_{k \neq p_0} Q_{p_0 k} \cdot \alpha'_{p_0 j} + \sum_{i \neq j, p_0} \sum_{k \neq i} Q_{ik} \cdot \alpha'_{ij} \right\} \\
&> \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ \sum_{k \neq p_0} Q_{p_0 k} \cdot \alpha'_{p_0 j} + \sum_{i \neq j, p_0} x^*_i \cdot \alpha'_{ij} \right\} \\
&= \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ \left(\sum_{k \neq p_0} Q_{p_0 k} - x^*_{p_0} \right) \cdot \alpha'_{p_0 j} + x^*_{p_0} \cdot \alpha'_{p_0 j} + \sum_{i \neq j, p_0} x^*_i \cdot \alpha'_{ij} \right\} \\
&= \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ \left(\sum_{k \neq p_0} Q_{p_0 k} - x^*_{p_0} \right) \cdot \alpha'_{p_0 j} + x^*_j \right\} \\
&\geq \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ \min_{v:\ v \neq j} \alpha'_{vj} \cdot \Big(\sum_{k=1}^{n} Q_{vk} - x^*_v \Big) + x^*_j \right\} \\
&= \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ \tau + x^*_j \right\} \\
&= \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk})
\end{aligned}
$$

Then we have

$$
\begin{aligned}
X'_j(2) &= \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ UI'_j(1) + \sum_i Y'_{ij}(1) \right\} \\
&\geq \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}) \right\} \\
&= \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk})
\end{aligned}
$$

So the relation (A.9) holds for $t = 2$ for the location j. For any location $v \neq j$, we have

$$\begin{aligned}
X'_v(2) =& \min\left\{\sum_{k=1}^{n} Q_{vk},\ \sum_{i\neq v} X'_i(1)\cdot\alpha'_{iv}\right\}\\
=& \min\left\{x^*_v,\ X'_j(1)\cdot\alpha_{jv} + \sum_{i\neq j,v} X'_i(1)\cdot\alpha_{iv}\right\}\\
\geq& \min\left\{x^*_v,\ (X'_j(1)-x^*_j)\cdot\alpha_{jv} + x^*_j\cdot\alpha_{jv} + \sum_{i\neq j,v} x^*_i\cdot\alpha_{iv}\right\}\\
=& \min\left\{x^*_v,\ (X'_j(1)-x^*_j)\cdot\alpha_{jv} + x^*_v\right\}\\
>& \min\left\{x^*_v,\ x^*_v\right\}\\
=& x^*_v
\end{aligned}$$

So the relation (A.9) holds for $t = 2$ for all locations.

Suppose the relation (A.9) holds for all periods up to the period $t \geq 2$ (including the period t). Next we prove it also holds for the period $t+1$.

We first prove that the relation (A.9) holds for the period $t+1$ for the location j. We need to prove $X'_j(t+1) \geq \sum_{k=1}^{n}(Q_{jk} + \tau\cdot\varepsilon_{jk})$. We discuss three cases:

Case 1: $UI'_j(t-1) + \sum_i Y'_{ij}(t-1) \geq 2\tau + \sum_{k=1}^{n} Q_{jk}$ holds. In this case, we must have

$$\begin{aligned}
UI'_j(t) =& \max\left\{0,\ UI'_j(t-1) + \sum_i Y'_{ij}(t-1) - \sum_{k=1}^{n} Q_{jk} - \tau\right\}\\
\geq& \tau
\end{aligned}$$

Then we have

$$\begin{aligned}
X'_j(t+1) &= \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ UI'_j(t)+\sum_i Y'_{ij}(t)\right\}\\
&= \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ UI'_j(t)+\sum_i X'_i(t)\cdot\alpha'_{ij}\right\}\\
&> \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ \tau+\sum_i x^*_i\cdot\alpha'_{ij}\right\}\\
&= \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ \tau+x^*_j\right\}\\
&= \sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk})
\end{aligned}$$

Case 2: $UI'_j(t-1)+\sum_i Y'_{ij}(t-1) < 2\tau+\sum_{k=1}^{n} Q_{jk}$ and there exists a location $m \neq j$ such that $UI'_m(t-1)+\sum_i Y'_{im}(t-1) \geq \sum_{k=1}^{n} Q_{mk}$. Then we have $X'_m(t) = \min\left\{\sum_{k=1}^{n} Q_{mk},\ UI'_m(t-1)+\sum_i Y'_{im}(t-1)\right\} \geq \sum_{k=1}^{n} Q_{mk}$. Further we have

$$\begin{aligned}
X'_j(t+1) &= \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ UI'_j(t)+\sum_i Y'_{ij}(t)\right\}\\
&= \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ UI'_j(t)+\sum_i X'_i(t)\cdot\alpha'_{ij}\right\}\\
&\geq \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ \sum_i X'_i(t)\cdot\alpha'_{ij}\right\}\\
&\geq \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ \left(\sum_{k=1}^{n} Q_{mk}\right)\cdot\alpha'_{mj}+\sum_{i\neq m} X'_i(t)\cdot\alpha'_{ij}\right\}\\
&\geq \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ \left(\sum_{k=1}^{n} Q_{mk}\right)\cdot\alpha'_{mj}+\sum_{i\neq m} x^*_i\cdot\alpha'_{ij}\right\}\\
&= \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ \left(-x^*_m+\sum_{k=1}^{n} Q_{mk}\right)\cdot\alpha'_{mj}+\sum_i x^*_i\cdot\alpha'_{ij}\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ \left(-x_m^* + \sum_{k=1}^{n} Q_{mk} \right) \cdot \alpha'_{mj} + x_j^* \right\} \\
&\geq \min\left\{ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}),\ \min_{v:\ v \neq j} \alpha'_{vj} \cdot \left(\sum_{k=1}^{n} Q_{vk} - x_v^* \right) + x_j^* \right\} \\
&= \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk})
\end{aligned}
$$

Case 3: $UI'_j(t-1) + \sum_i Y'_{ij}(t-1) < 2\tau + \sum_{k=1}^{n} Q_{jk}$ and $UI'_v(t-1) + \sum_i Y'_{iv}(t-1) < \sum_{k=1}^{n} Q_{vk}, \forall v \neq j$.

Because for $v \neq j$, we have $x_v^* < X'_v(t) = \min\{\sum_{k=1}^{n} Q_{vk},\ UI'_v(t-1) + \sum_i Y'_{iv}(t-1)\} = UI'_v(t-1) + \sum_i Y'_{iv}(t-1)$, we can denote $UI'_v(t-1) + \sum_i Y'_{iv}(t-1) = x_v^* + \phi_v$, where ϕ_v satisfies $0 < \phi_v < \sum_{k=1}^{n} Q_{vk} - x_v^*$. Note that the total number of vehicles in the network is $\sum_k UI'_k(t-1) + \sum_k \sum_i Y'_{ik}(t-1) = \tau + \sum_i \sum_k Q_{ik}$. We have

$$
\begin{aligned}
&\sum_{v \neq j} \phi_v \\
&= \sum_{v \neq j} UI'_v(t-1) + \sum_{v \neq j} \sum_i Y'_{iv}(t-1) - \sum_{v \neq j} x_v^* \\
&= \sum_k UI'_k(t-1) + \sum_k \sum_i Y'_{ik}(t-1) - \left[UI'_j(t-1) + \sum_i Y'_{ij}(t-1) \right] - \sum_{v \neq j} x_v^* \\
&> \sum_{i \neq j} \sum_k Q_{ik} + \tau - (2\tau + \sum_{k=1}^{n} Q_{ik}) - \sum_{v \neq j} x_v^* \\
&= \sum_{v \neq j} \sum_k Q_{vk} - \sum_{v \neq j} x_v^* - \tau
\end{aligned}
$$

Suppose $m_0 = \arg\max_{v \neq j} \left(\sum_k Q_{vk} - x_v^* \right)$. Because the network has at least 3 locations, there must be a location m_1 such that $m_1 \neq m_0,\ m_1 \neq j$. Then we have

$$
\sum_{v \neq j} \phi_v \tag{A.10a}
$$

$$> \sum_{v\neq j}\sum_{k} Q_{vk} - \sum_{v\neq j} x^*_v - \tau \tag{A.10b}$$

$$= \sum_{k} Q_{m_0 k} - x^*_{m_0} + \sum_{k} Q_{m_1 k} - x^*_{m_1} + \sum_{v\neq j, m_0, m_1}\sum_{k} Q_{vk} - \sum_{v\neq j, m_0, m_1} x^*_v - \tau \tag{A.10c}$$

$$\geq \sum_{k} Q_{m_0 k} - x^*_{m_0} + \sum_{k} Q_{m_1 k} - x^*_{m_1} - \tau \tag{A.10d}$$

$$\geq \sum_{k} Q_{m_0 k} - x^*_{m_0} + \min_{v:\ v\neq j} \alpha'_{vj} \cdot \left(\sum_{k=1}^{n} Q_{vk} - x^*_v \right) - \tau \tag{A.10e}$$

$$\geq \sum_{k} Q_{m_0 k} - x^*_{m_0} \tag{A.10f}$$

$$= \max_{v\neq j} \left(\sum_{k} Q_{vk} - x^*_v \right) \tag{A.10g}$$

Further, we have

$$\begin{aligned}
&\sum_{v\neq j} \phi_v \cdot \alpha_{vj} \\
&> \frac{\max_{p\neq j} \left(\sum_k Q_{pk} - x^*_p \right)}{\sum_{m\neq j} \phi_m} \cdot \sum_{v\neq j} \phi_v \cdot \alpha_{vj} \\
&= \sum_{v\neq j} \left[\frac{\phi_v}{\sum_{m\neq j} \phi_m} \cdot \alpha_{vj} \cdot \max_{p\neq j} \left(\sum_k Q_{pk} - x^*_p \right) \right] \\
&\geq \sum_{v\neq j} \left[\frac{\phi_v}{\sum_{m\neq j} \phi_m} \cdot \alpha_{vj} \cdot \left(\sum_k Q_{vk} - x^*_v \right) \right] \\
&\geq \sum_{v\neq j} \left[\frac{\phi_v}{\sum_{m\neq j} \phi_m} \cdot \min_{p:\ p\neq j} \alpha'_{pj} \cdot \left(\sum_{k=1}^{n} Q_{pk} - x^*_p \right) \right] \\
&= \min_{v:\ v\neq j} \alpha'_{vj} \cdot \left(\sum_{k=1}^{n} Q_{vk} - x^*_v \right) \\
&\geq \tau
\end{aligned}$$

where the first inequality is due to (A.10).

Then we have

$$
\begin{aligned}
&\sum_{v\neq j} X'_v(t)\cdot\alpha'_{vj} \\
=&\sum_{v\neq j}\left[\min\left\{\sum_{k=1}^{n} Q_{vk},\ UI'_v(t-1)+\sum_i Y'_{iv}(t-1)\right\}\cdot\alpha'_{vj}\right] \\
=&\sum_{v\neq j}\left[UI'_v(t-1)+\sum_i Y'_{iv}(t-1)\right]\cdot\alpha'_{vj} \\
=&\sum_{v\neq j}(x^*_v+\phi_v)\cdot\alpha'_{vj} \\
=&x^*_j+\sum_{v\neq j}\phi_v\cdot\alpha'_{vj} \\
\geq&x^*_j+\tau
\end{aligned}
$$

Then we have

$$
\begin{aligned}
X'_j(t+1) =& \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ UI'_j(t)+\sum_i Y'_{ij}(t)\right\} \\
=& \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ UI'_j(t)+\sum_i X'_i(t)\cdot\alpha'_{ij}\right\} \\
\geq& \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ \sum_i X'_i(t)\cdot\alpha'_{ij}\right\} \\
\geq& \min\left\{\sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk}),\ x^*_j+\tau\right\} \\
=& \sum_{k=1}^{n}(Q_{jk}+\tau\cdot\varepsilon_{jk})
\end{aligned}
$$

Therefore, we have proven that the relation (A.9) holds for the period $t+1$ for the location j.

We next prove that the relation (A.9) holds for the period $t+1$ for any location $v \neq j$. For any location $v \neq j$, we have

$$
\begin{aligned}
X'_v(t+1) &= \min\left\{ \sum_{k=1}^{n} Q_{vk},\ UI'_v(t) + \sum_i Y'_{iv}(t) \right\} \\
&\geq \min\left\{ x^*_v,\ \sum_i X'_i(t) \cdot \alpha_{iv} \right\} \\
&\geq \min\left\{ x^*_v,\ X'_j(t) \cdot \alpha'_{jv} + \sum_{i \neq j} X'_i(t) \cdot \alpha'_{iv} \right\} \\
&\geq \min\left\{ x^*_v,\ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}) \cdot \alpha'_{jv} + \sum_{i \neq j} x^*_i \cdot \alpha'_{iv} \right\} \\
&= \min\left\{ x^*_v,\ \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk}) \cdot \frac{Q_{jv} + \tau \cdot \varepsilon_{jv}}{\sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk})} + \sum_{i \neq j} x^*_i \cdot \alpha'_{iv} \right\} \\
&= \min\left\{ x^*_v,\ Q_{jv} + \tau \cdot \varepsilon_{jv} + \sum_{i \neq j} x^*_i \cdot \alpha'_{iv} \right\} \\
&\geq \min\left\{ x^*_v,\ \tau \cdot \varepsilon_{jv} + \sum_i x^*_i \cdot \alpha'_{iv} \right\} \\
&\geq \min\left\{ x^*_v,\ \sum_i x^*_i \cdot \alpha'_{iv} \right\} \\
&= \min\left\{ x^*_v,\ x^*_v \right\} \\
&= x^*_v
\end{aligned}
$$

Therefore, we have proven that the relation (A.9) holds for the period $t+1$ for any location in the network.

Finally, we prove if $0 < \tau \leq \min_{v:\ v \neq j} \alpha'_{vj} \cdot \left(\sum_{k=1}^{n} Q_{vk} - x^*_v \right)$, then $\widetilde{x}^*_j = \tau + \sum_{k=1}^{n} Q_{jk}$ and $\widetilde{x}^*_v < \sum_{k=1}^{n} Q_{vk}, \forall v \neq j$ hold. We have $\widetilde{x}^*_j = \lim_{t \to \infty} X'_j(t) \geq \sum_{k=1}^{n} (Q_{jk} + \tau \cdot \varepsilon_{jk})$. Because we also have

$\widetilde{x}_j^* = \lim_{t\to\infty} X_j'(t) \le \sum_{k=1}^n (Q_{jk} + \tau \cdot \varepsilon_{jk})$, we have that $\widetilde{x}_j^* = \sum_{k=1}^n (Q_{jk} + \tau \cdot \varepsilon_{jk}) = \tau + \sum_{k=1}^n Q_{jk}$. We also have that $\widetilde{x}_v^* = \lim_{t\to\infty} X_v'(t) > x_v^*, \forall v \neq j$.

We then prove that $\widetilde{x}_v^* < \sum_{k=1}^n Q_{vk}, \forall v \neq j$ hold. Suppose not, then there exists a location $r_0 \neq j$ such that $\widetilde{x}_{r_0}^* = \sum_{k=1}^n Q_{r_0 k}$. Then

$$
\begin{aligned}
\widetilde{x}_j^* &= \sum_{i\neq j} \widetilde{x}_i^* \cdot \alpha_{ij}' \\
&= \widetilde{x}_{r_0}^* \cdot \alpha_{r_0 j}' + \sum_{i\neq j, r_0} \widetilde{x}_i^* \cdot \alpha_{ij}' \\
&= \sum_{k=1}^n Q_{r_0 k} \cdot \alpha_{r_0 j}' + \sum_{i\neq j, r_0} \widetilde{x}_i^* \cdot \alpha_{ij}' \\
&> \sum_{k=1}^n Q_{r_0 k} \cdot \alpha_{r_0 j}' + \sum_{i\neq j, r_0} x_i^* \cdot \alpha_{ij}' \\
&= \left(\sum_{k=1}^n Q_{r_0 k} - x_{r_0}^* \right) \cdot \alpha_{r_0 j}' + \sum_{i\neq j} x_i^* \cdot \alpha_{ij}' \\
&\ge \min_{v:\ v\neq j} \Big(\sum_{k=1}^n Q_{vk} - x_v^* \Big) \cdot \alpha_{vj}' + \sum_{i\neq j} x_i^* \cdot \alpha_{ij}' \\
&\ge \tau + x_j^* \\
&= \tau + \sum_{k=1}^n Q_{jk}
\end{aligned}
$$

So we have $\widetilde{x}_j^* > \tau + \sum_{k=1}^n Q_{jk}$, which contradicts that $\widetilde{x}_j^* \le \tau + \sum_{k=1}^n Q_{jk}$. Hence proven. □

A.7 Proof of Proposition 7.

We consider the following two cases.

1. Scenario 1: at the beginning of each period, there are Q_{mv} potential customers seeking to go from location m to location v. At the beginning of $t = 1$, we place

$\sum_{m=1}^{n} Q_{vm}$ vehicles at each location $v \in \{1, \cdots, n\}$. Denote the flow from location m to location v during the time period t as $Y_{mv}(t)$. Denote the outbound flow from location v during time period t as $X_v(t)$. We use $UI_v(t)$ to denote the number of vehicles that are not used by any customer during the period t and stay at location v during the period t.

2. Scenario 2: at the beginning of each period, there are Q_{vm} potential customers seeking to go from location v to location m (except from location i to location k), and $Q_{ik} + \tau$ potential customers seeking to go from location i to location k. At the beginning of $t = 1$, we place $\sum_{v=1}^{n} Q_{mv}$ vehicles at each location $v \in \{1, \cdots, n\} \setminus \{i\}$, and $\sum_{m=1}^{n} Q_{im} + \tau$ vehicles at the location i. Denote the flow from location m to location v during the time period t as $Y'_{mv}(t)$. Denote the outbound flow from location v during time period t as $X'_v(t)$. We use $UI'_v(t)$ to denote the number of vehicles that are not used by any customer during the period t and stay at location v during the period t.

We just need to prove if τ is no greater than the maximum $\widehat{B}_{ik}$ value that satisfies the corresponding inequalities, then the location j remains to be the only critical location in Scenario 2. We prove this by showing that in Scenario 2: (A) the location i cannot be the critical location; (B) the location k cannot be the critical location; (C) any location q ($q \neq i, j, k$) cannot be the critical location.

In Scenario 1, for each arc (m, v), denote the associated fraction $\frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}}$ as α_{mv}. In Scenario 2, for each arc (m, v) such that $m \neq i$, denote the associated fraction $\frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}}$ as α'_{mv}; for the arc (i, m) such that $m \neq k$, denote the associated fraction $\frac{Q_{im}}{\tau + \sum_{l=1}^{n} Q_{il}}$ as α'_{im}; for the arc (i, k), denote the associated fraction $\frac{Q_{ik}+\tau}{\tau+\sum_{l=1}^{n} Q_{il}}$ as α'_{ik}.

According to Proposition 2, we know the system state in each of the above two scenarios will

evolve into the corresponding equilibrium flow. Denote the equilibrium flow in Scenario 1 and Scenario 2 as $\boldsymbol{x}^*$ and $\widetilde{\boldsymbol{x}}^*$. According to Lemma 2, we know $X_v(t) \geq x_v^*$ and $X_v'(t) \geq \widetilde{x}_v^*$ $\forall v = 1, \cdots, n$, $\forall t \geq 1$.

Part A. We prove that the location i cannot be the critical location in Scenario 2. Suppose not, and the location i is the critical location in Scenario 2. Then we have $\widetilde{x}_i^* = \tau + \sum_{l=1}^n Q_{il}$. According to Lemma 2, we have $X_i'(t) \geq \widetilde{x}_i^*$, $\forall t \geq 1$. Because we know $X_i'(t) \leq \tau + \sum_{l=1}^n Q_{il} = \widetilde{x}_i^*$, $\forall t \geq 1$, we have that $X_i'(t) = \widetilde{x}_i^*$, $\forall t \geq 1$.

We next show that the following relation holds.

$$X_v'(t) \geq X_v(t),\ \forall v = 1, \cdots, n,\ \forall t \geq 1 \tag{A.11a}$$

$$UI_v'(t) \geq UI_v(t),\ \forall v = 1, \cdots, n,\ \forall t \geq 1 \tag{A.11b}$$

It is straightforward to see that the relation (A.11) holds for $t = 1$. Suppose the relation (A.11) holds for the time period t, we next show that it also holds for the time period $t+1$.

First for the location i, for all $t \geq 1$, we have that $X_i'(t) = \widetilde{x}_i^* = \tau + \sum_{l=1}^n Q_{il} \geq X_i(t)$.

$$\begin{aligned}
UI_i'(t+1) =& \max\left\{0,\ UI_i'(t) + \sum_{v \neq i} Y_{vi}'(t) - \tau - \sum_{m=1}^n Q_{im}\right\} \\
=& \max\left\{0,\ UI_i'(t) + \sum_{v \neq i} X_v'(t) \cdot \frac{Q_{vi}}{\sum_{m=1}^n Q_{vm}} - \tau - \sum_{m=1}^n Q_{im}\right\} \\
\geq& \max\left\{0,\ UI_i(t) + \sum_{v \neq i} X_v(t) \cdot \frac{Q_{vi}}{\sum_{m=1}^n Q_{vm}} - \sum_{m=1}^n Q_{im}\right\} \\
=& UI_i(t+1)
\end{aligned}$$

So the relation (A.11) holds for the location i for the time period $t+1$.

For the location k, we have

$$
\begin{aligned}
X'_k(t+1) &= \min\left\{ \sum_{m=1}^{n} Q_{km},\ UI'_k(t) + \sum_{m} Y'_{mk}(t) \right\} \\
&= \min\left\{ \sum_{m=1}^{n} Q_{km},\ UI'_k(t) + Y'_{ik}(t) + \sum_{m\neq i} Y'_{mk}(t) \right\} \\
&= \min\left\{ \sum_{m=1}^{n} Q_{km},\ UI'_k(t) + \tau + Q_{ik} + \sum_{m\neq i} Y'_{mk}(t) \right\} \\
&= \min\left\{ \sum_{m=1}^{n} Q_{km},\ UI'_k(t) + \tau + Q_{ik} + \sum_{m\neq i} X'_m(t) \cdot \frac{Q_{mk}}{\sum_{l=1}^{n} Q_{ml}} \right\} \\
&> \min\left\{ \sum_{m=1}^{n} Q_{km},\ UI_k(t) + Q_{ik} + \sum_{m\neq i} X_m(t) \cdot \frac{Q_{mk}}{\sum_{l=1}^{n} Q_{ml}} \right\} \\
&\geq \min\left\{ \sum_{m=1}^{n} Q_{km},\ UI_k(t) + X_i(t) \cdot \frac{Q_{ik}}{\sum_{l=1}^{n} Q_{il}} + \sum_{m\neq i} X_m(t) \cdot \frac{Q_{mk}}{\sum_{l=1}^{n} Q_{ml}} \right\} \\
&= \min\left\{ \sum_{m=1}^{n} Q_{km},\ UI_k(t) + \sum_{m} X_m(t) \cdot \frac{Q_{mk}}{\sum_{l=1}^{n} Q_{ml}} \right\} \\
&= X_k(t+1),
\end{aligned}
$$

and

$$
\begin{aligned}
UI'_k(t+1) &= \max\left\{ 0,\ UI'_k(t) + \sum_{v} Y'_{vk}(t) - \sum_{m=1}^{n} Q_{km} \right\} \\
&= \max\left\{ 0,\ UI'_k(t) + Y'_{ik}(t) + \sum_{m\neq i} Y'_{mk}(t) - \sum_{m=1}^{n} Q_{km} \right\} \\
&= \max\left\{ 0,\ UI'_k(t) + \tau + Q_{ik} + \sum_{m\neq i} X'_m(t) \cdot \frac{Q_{mk}}{\sum_{l=1}^{n} Q_{ml}} - \sum_{m=1}^{n} Q_{km} \right\} \\
&> \max\left\{ 0,\ UI'_k(t) + Q_{ik} + \sum_{m\neq i} X_m(t) \cdot \frac{Q_{mk}}{\sum_{l=1}^{n} Q_{ml}} - \sum_{m=1}^{n} Q_{km} \right\}
\end{aligned}
$$

$$
\begin{aligned}
&\geq \max\left\{0,\ UI'_k(t)+X_i(t)\cdot\frac{Q_{ik}}{\sum_{l=1}^n Q_{il}}+\sum_{m\neq i}X_m(t)\cdot\frac{Q_{mk}}{\sum_{l=1}^n Q_{ml}}-\sum_{m=1}^n Q_{km}\right\}\\
&=\max\left\{0,\ UI'_k(t)+\sum_{m}X_m(t)\cdot\frac{Q_{mk}}{\sum_{l=1}^n Q_{ml}}-\sum_{m=1}^n Q_{km}\right\}\\
&\geq \max\left\{0,\ UI_k(t)+\sum_{m}X_m(t)\cdot\frac{Q_{mk}}{\sum_{l=1}^n Q_{ml}}-\sum_{m=1}^n Q_{km}\right\}\\
&=UI_k(t+1)
\end{aligned}
$$

So the relation (A.11) holds for the location k for the time period $t+1$.

For any location $v\neq i,k$, we have

$$
\begin{aligned}
X'_v(t+1) &= \min\left\{\sum_{m=1}^n Q_{vm},\ UI'_v(t)+\sum_{m}Y'_{mv}(t)\right\}\\
&= \min\left\{\sum_{m=1}^n Q_{vm},\ UI'_v(t)+Y'_{iv}(t)+\sum_{m\neq i}Y'_{mv}(t)\right\}\\
&= \min\left\{\sum_{m=1}^n Q_{vm},\ UI'_v(t)+Q_{iv}+\sum_{m\neq i}Y'_{mv}(t)\right\}\\
&= \min\left\{\sum_{m=1}^n Q_{vm},\ UI'_v(t)+Q_{iv}+\sum_{m\neq i}X'_m(t)\cdot\frac{Q_{mv}}{\sum_{l=1}^n Q_{ml}}\right\}\\
&\geq \min\left\{\sum_{m=1}^n Q_{vm},\ UI_v(t)+Q_{iv}+\sum_{m\neq i}X_m(t)\cdot\frac{Q_{mv}}{\sum_{l=1}^n Q_{ml}}\right\}\\
&\geq \min\left\{\sum_{m=1}^n Q_{vm},\ UI_v(t)+X_i(t)\cdot\frac{Q_{iv}}{\sum_{l=1}^n Q_{il}}+\sum_{m\neq i}X_m(t)\cdot\frac{Q_{mv}}{\sum_{l=1}^n Q_{ml}}\right\}\\
&= \min\left\{\sum_{m=1}^n Q_{vm},\ UI_v(t)+\sum_{m}X_m(t)\cdot\frac{Q_{mv}}{\sum_{l=1}^n Q_{ml}}\right\}\\
&=X_v(t+1),
\end{aligned}
$$

and

$$
\begin{aligned}
UI'_v(t+1) =& \max\left\{0,\ UI'_v(t)+\sum_m Y'_{mv}(t)-\sum_{m=1}^{n} Q_{vm}\right\}\\
=& \max\left\{0,\ UI'_v(t)+Y'_{iv}(t)+\sum_{m\neq i} Y'_{mv}(t)-\sum_{m=1}^{n} Q_{vm}\right\}\\
=& \max\left\{0,\ UI'_v(t)+Q_{iv}+\sum_{m\neq i} X'_m(t)\cdot\frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}}-\sum_{m=1}^{n} Q_{vm}\right\}\\
\geq& \max\left\{0,\ UI_v(t)+Q_{iv}+\sum_{m\neq i} X_m(t)\cdot\frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}}-\sum_{m=1}^{n} Q_{vm}\right\}\\
\geq& \max\left\{0,\ UI_v(t)+X_i(t)\cdot\frac{Q_{iv}}{\sum_{l=1}^{n} Q_{il}}+\sum_{m\neq i} X_m(t)\cdot\frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}}-\sum_{m=1}^{n} Q_{vm}\right\}\\
=& \max\left\{0,\ UI_v(t)+\sum_m X_m(t)\cdot\frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}}-\sum_{m=1}^{n} Q_{vm}\right\}\\
=& UI_v(t+1)
\end{aligned}
$$

Therefore we have proven that the relation (A.11) holds for all locations for the time period $t+1$. As a result, we have $\widetilde{x}_v^* = \lim_{t\to\infty} X'_v(t) \geq \lim_{t\to\infty} X_v(t) = x_v^*,\ \forall v = 1,\cdots,n$. Because we have $\widetilde{x}_i^* = \tau + \sum_{l=1}^{n} Q_{il}$ and $x_i^* < \sum_{l=1}^{n} Q_{il}$, we have

$$
\begin{aligned}
\widetilde{x}_j^* =& \sum_{v=1}^{n} \widetilde{x}_v^* \cdot \alpha'_{vj}\\
\geq& \widetilde{x}_i^* \cdot \alpha'_{ij} + \sum_{v\neq i} \widetilde{x}_v^* \cdot \alpha'_{vj}\\
=& \widetilde{x}_i^* \cdot \frac{Q_{ij}}{\tau + \sum_{l=1}^{n} Q_{il}} + \sum_{v\neq i} \widetilde{x}_v^* \cdot \alpha'_{vj}\\
=& Q_{ij} + \sum_{v\neq i} \widetilde{x}_v^* \cdot \alpha'_{vj}
\end{aligned}
$$

$$
\begin{aligned}
&> x_i^* \cdot \frac{Q_{ij}}{\sum_{l=1}^n Q_{il}} + \sum_{v \neq i} \widetilde{x}_v^* \cdot \alpha'_{vj} \\
&= x_i^* \cdot \frac{Q_{ij}}{\sum_{l=1}^n Q_{il}} + \sum_{v \neq i} \widetilde{x}_v^* \cdot \alpha_{vj} \\
&\geq x_i^* \cdot \frac{Q_{ij}}{\sum_{l=1}^n Q_{il}} + \sum_{v \neq i} x_v^* \cdot \alpha_{vj} \\
&= \sum_v x_v^* \cdot \alpha_{vj} \\
&= x_j^* \\
&= \sum_{l=1}^n Q_{jl},
\end{aligned}
$$

which contradicts that $\widetilde{x}_j^* \leq \sum_{l=1}^n Q_{jl}$. Therefore, the location i cannot be the critical location in Scenario 2.

Part B. We prove that the location k cannot be the critical location in Scenario 2. Suppose not, and the location k is the critical location in Scenario 2. Then we have $\widetilde{x}_k^* = \sum_{l=1}^n Q_{kl}$. According to Lemma 2, we have $X'_k(t) \geq \widetilde{x}_k^*,\ \forall t \geq 1$. Because we know $X'_k(t) \leq \sum_{l=1}^n Q_{kl} = \widetilde{x}_k^*,\ \forall t \geq 1$, we have that $X'_k(t) = \widetilde{x}_k^* = \sum_{l=1}^n Q_{kl},\ \forall t \geq 1$.

We next show that the following relation holds.

$$
X'_v(t) \geq x_v^*,\ \forall v \neq i,k,\ \forall t \geq 1 \tag{A.12a}
$$

$$
X'_i(t) \geq \min \left\{ \sum_{l=1}^n Q_{il},\ x_i^* + \left[\sum_{m=1}^n Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\},\ \forall t \geq 1 \tag{A.12b}
$$

$$
X'_k(t) = \sum_{l=1}^n Q_{kl},\ \forall t \geq 1 \tag{A.12c}
$$

We already proven the last relation. We just need to prove the first two relations. It is straightforward to see that the relation (A.12) holds for $t = 1$. Suppose the relation (A.12) holds for the time

period t, we next show that it also holds for the time period $t+1$.

For the location i, we have

$$
\begin{aligned}
& X_i'(t+1) \\
&= \min\left\{ \sum_{m=1}^{n} Q_{im},\ UI_i'(t) + \sum_{m} Y_{mi}'(t) \right\} \\
&\geq \min\left\{ \sum_{m=1}^{n} Q_{im},\ Y_{ki}'(t) + \sum_{m\neq i,k} Y_{mi}'(t) \right\} \\
&= \min\left\{ \sum_{m=1}^{n} Q_{im},\ X_k'(t)\cdot \frac{Q_{ki}}{\sum_{l=1}^{n} Q_{kl}} + \sum_{m\neq i,k} X_m'(t)\cdot \frac{Q_{mi}}{\sum_{l=1}^{n} Q_{ml}} \right\} \\
&= \min\left\{ \sum_{m=1}^{n} Q_{im},\ \frac{Q_{ki}}{\sum_{l=1}^{n} Q_{kl}} \cdot \sum_{l=1}^{n} Q_{kl} + \sum_{m\neq i,k} x_m^* \cdot \frac{Q_{mi}}{\sum_{l=1}^{n} Q_{ml}} \right\} \\
&= \min\left\{ \sum_{m=1}^{n} Q_{im},\ \left(\sum_{l=1}^{n} Q_{kl} - x_k^* \right) \cdot \frac{Q_{ki}}{\sum_{l=1}^{n} Q_{kl}} + \sum_{m\neq i} x_m^* \cdot \frac{Q_{mi}}{\sum_{l=1}^{n} Q_{ml}} \right\} \\
&= \min\left\{ \sum_{m=1}^{n} Q_{im},\ \left(\sum_{l=1}^{n} Q_{kl} - x_k^* \right) \cdot \alpha_{ki} + x_i^* \right\}
\end{aligned}
$$

So the relation (A.12) holds for the location i for the time period $t+1$. For any location $v \neq i,k$, we have

$$
\begin{aligned}
& X_v'(t+1) \\
&= \min\left\{ \sum_{m=1}^{n} Q_{vm},\ UI_v'(t) + \sum_{m} Y_{mv}'(t) \right\} \\
&\geq \min\left\{ \sum_{m=1}^{n} Q_{vm},\ Y_{iv}'(t) + Y_{kv}'(t) + \sum_{m\neq i,k} Y_{mv}'(t) \right\} \\
&= \min\left\{ \sum_{m=1}^{n} Q_{vm},\ X_i'(t)\cdot \frac{Q_{iv}}{\tau + \sum_{l=1}^{n} Q_{il}} + X_k'(t)\cdot \frac{Q_{kv}}{\sum_{l=1}^{n} Q_{kl}} + \sum_{m\neq i,k} X_m'(t)\cdot \frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}} \right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \min\left\{ \sum_{m=1}^{n} Q_{vm},\ X_i'(t)\cdot\frac{Q_{iv}}{\tau+\sum_{l=1}^{n}Q_{il}} + \sum_{m=1}^{n} Q_{km}\cdot\frac{Q_{kv}}{\sum_{l=1}^{n}Q_{kl}} + \sum_{m\neq i,k} X_m'(t)\cdot\frac{Q_{mv}}{\sum_{l=1}^{n}Q_{ml}} \right\} \\
&\geq \min\left\{ \sum_{m=1}^{n} Q_{vm},\ \min\left\{ \sum_{l=1}^{n} Q_{il},\ x_i^* + \left[\sum_{m=1}^{n} Q_{km} - x_k^*\right]\cdot\alpha_{ki} \right\}\cdot\frac{Q_{iv}}{\tau+\sum_{l=1}^{n}Q_{il}} \right. \\
&\qquad\qquad \left. + \sum_{m=1}^{n} Q_{km}\cdot\frac{Q_{kv}}{\sum_{l=1}^{n}Q_{kl}} + \sum_{m\neq i,k} x_m^*\cdot\frac{Q_{mv}}{\sum_{l=1}^{n}Q_{ml}} \right\} \\
&= \min\left\{ \sum_{m=1}^{n} Q_{vm},\ x_i^*\cdot\left[\frac{Q_{iv}}{\tau+\sum_{l=1}^{n}Q_{il}} - \frac{Q_{iv}}{\sum_{l=1}^{n}Q_{il}}\right] \right. \\
&\qquad\qquad + \min\left\{ \sum_{l=1}^{n} Q_{il} - x_i^*,\ \left[\sum_{m=1}^{n} Q_{km} - x_k^*\right]\cdot\alpha_{ki} \right\}\cdot\frac{Q_{iv}}{\tau+\sum_{l=1}^{n}Q_{il}} \\
&\qquad\qquad \left. + \left[\sum_{m=1}^{n} Q_{km} - x_k^*\right]\cdot\frac{Q_{kv}}{\sum_{l=1}^{n}Q_{kl}} + \sum_{m} x_m^*\cdot\frac{Q_{mv}}{\sum_{l=1}^{n}Q_{ml}} \right\} \\
&= \min\left\{ \sum_{m=1}^{n} Q_{vm},\ x_v^* + x_i^*\cdot\left[\frac{Q_{iv}}{\tau+\sum_{l=1}^{n}Q_{il}} - \frac{Q_{iv}}{\sum_{l=1}^{n}Q_{il}}\right] \right. \\
&\qquad\qquad + \min\left\{ \sum_{l=1}^{n} Q_{il} - x_i^*,\ \left[\sum_{m=1}^{n} Q_{km} - x_k^*\right]\cdot\alpha_{ki} \right\}\cdot\frac{Q_{iv}}{\tau+\sum_{l=1}^{n}Q_{il}} \\
&\qquad\qquad \left. + \left[\sum_{m=1}^{n} Q_{km} - x_k^*\right]\cdot\frac{Q_{kv}}{\sum_{l=1}^{n}Q_{kl}} \right\} \\
&\overset{(B1)}{\geq} \min\left\{ \sum_{m=1}^{n} Q_{vm},\ x_v^* \right\} \\
&= x_v^*
\end{aligned}
$$

where the inequality $(B1)$ is due to the following relation

$$
\tau \leq \frac{\sum_{l=1}^{n} Q_{il}}{x_i^*}\cdot\min\left\{ \sum_{l=1}^{n} Q_{il} - x_i^*,\ \left[\sum_{m=1}^{n} Q_{km} - x_k^*\right]\cdot\alpha_{ki} \right\} \tag{A.13a}
$$

$$
+ \frac{\sum_{l=1}^{n} Q_{il}}{x_i^*}\cdot\frac{\alpha_{kv}}{\alpha_{iv}}\cdot\left[\sum_{m=1}^{n} Q_{km} - x_k^*\right],\ \forall v \neq i,k \tag{A.13b}
$$

$$\implies x_i^* \cdot \alpha_{iv} \cdot \tau \leq \min\left\{ \sum_{l=1}^{n} Q_{il} - x_i^*,\ \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\} \cdot Q_{iv} \tag{A.13c}$$

$$+ \left[\sum_{l=1}^{n} Q_{il} \right] \cdot \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{kv},\ \forall v \neq i,k \tag{A.13d}$$

$$\implies \frac{1}{\sum_{l=1}^{n} Q_{il}} \cdot \left(x_i^* \cdot \alpha_{iv} \cdot \tau - \min\left\{ \sum_{l=1}^{n} Q_{il} - x_i^*,\ \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\} \cdot Q_{iv} \right) \tag{A.13e}$$

$$\leq \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{kv},\ \forall v \neq i,k \tag{A.13f}$$

$$\implies \frac{1}{\tau + \sum_{l=1}^{n} Q_{il}} \cdot \left(x_i^* \cdot \alpha_{iv} \cdot \tau - \min\left\{ \sum_{l=1}^{n} Q_{il} - x_i^*,\ \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\} \cdot Q_{iv} \right) \tag{A.13g}$$

$$< \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{kv},\ \forall v \neq i,k \tag{A.13h}$$

$$\implies x_i^* \cdot \alpha_{iv} \cdot \frac{\tau}{\tau + \sum_{l=1}^{n} Q_{il}} < \min\left\{ \sum_{l=1}^{n} Q_{il} - x_i^*,\ \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\} \cdot \frac{Q_{iv}}{\tau + \sum_{l=1}^{n} Q_{il}} \tag{A.13i}$$

$$+ \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{kv},\ \forall v \neq i,k \tag{A.13j}$$

$$\implies x_i^* \cdot \left[\frac{\tau \cdot Q_{iv}}{(\sum_{l=1}^{n} Q_{il}) \cdot (\tau + \sum_{l=1}^{n} Q_{il})} \right] < \tag{A.13k}$$

$$\min\left\{ \sum_{l=1}^{n} Q_{il} - x_i^*,\ \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\} \cdot \frac{Q_{iv}}{\tau + \sum_{l=1}^{n} Q_{il}} \tag{A.13l}$$

$$+ \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{kv},\ \forall v \neq i,k \tag{A.13m}$$

$$\implies x_i^* \cdot \left[\frac{Q_{iv}}{\sum_{l=1}^{n} Q_{il}} - \frac{Q_{iv}}{\tau + \sum_{l=1}^{n} Q_{il}} \right] < \tag{A.13n}$$

$$\min\left\{ \sum_{l=1}^{n} Q_{il} - x_i^*,\ \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\} \cdot \frac{Q_{iv}}{\tau + \sum_{l=1}^{n} Q_{il}} \tag{A.13o}$$

$$+ \left[\sum_{m=1}^{n} Q_{km} - x_k^* \right] \cdot \alpha_{kv},\ \forall v \neq i,k \tag{A.13p}$$

$$\Longrightarrow x_i^* \cdot \left[\frac{Q_{iv}}{\tau + \sum_{l=1}^n Q_{il}} - \frac{Q_{iv}}{\sum_{l=1}^n Q_{il}} \right] \tag{A.13q}$$

$$+ \min \left\{ \sum_{l=1}^n Q_{il} - x_i^*,\ \left[\sum_{m=1}^n Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\} \cdot \frac{Q_{iv}}{\tau + \sum_{l=1}^n Q_{il}} \tag{A.13r}$$

$$+ \left[\sum_{m=1}^n Q_{km} - x_k^* \right] \cdot \alpha_{kv} > 0,\ \forall v \neq i,k \tag{A.13s}$$

So the relation (A.12) holds for the time period $t+1$ for any location $v \neq i,k$.

Therefore, the relation (A.12) holds. Further, we have

$$\begin{aligned}
\widetilde{x}_v^* &= \lim_{t\to\infty} X_v'(t) \geq x_v^*,\ \forall v \neq i,k, \\
\widetilde{x}_i^* &= \lim_{t\to\infty} X_i'(t) \geq \min \left\{ \sum_{l=1}^n Q_{il},\ x_i^* + \left[\sum_{m=1}^n Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\}, \\
\widetilde{x}_k^* &= \lim_{t\to\infty} X_k'(t) = \sum_{l=1}^n Q_{kl},
\end{aligned}$$

However, we have

$$\begin{aligned}
\widetilde{x}_j^* &= \sum_{m\neq j} \widetilde{x}_m^* \cdot \alpha_{mj}' \\
&= \widetilde{x}_i^* \cdot \alpha_{ij}' + \widetilde{x}_k^* \cdot \alpha_{kj}' + \sum_{m\neq i,k} \widetilde{x}_m^* \cdot \alpha_{mj}' \\
&\geq \min \left\{ \sum_{l=1}^n Q_{il},\ x_i^* + \left[\sum_{m=1}^n Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\} \cdot \alpha_{ij}' + \sum_{l=1}^n Q_{kl} \cdot \alpha_{kj}' + \sum_{m\neq i,k} x_m^* \cdot \alpha_{mj}' \\
&= x_i^* \cdot \left[\frac{Q_{ij}}{\tau + \sum_{l=1}^n Q_{il}} - \frac{Q_{ij}}{\sum_{l=1}^n Q_{il}} \right] \\
&\qquad + \min \left\{ \sum_{l=1}^n Q_{il} - x_i^*,\ \left[\sum_{m=1}^n Q_{km} - x_k^* \right] \cdot \alpha_{ki} \right\} \cdot \frac{Q_{ij}}{\tau + \sum_{l=1}^n Q_{il}} \\
&\qquad + \left[\sum_{m=1}^n Q_{km} - x_k^* \right] \cdot \alpha_{kj} + \sum_m x_m^* \cdot \alpha_{mj}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(B2)}{>} \sum_m x_m^* \cdot \alpha_{mj} \\
&= x_j^* \\
&= \sum_{l=1}^{n} Q_{jl},
\end{aligned}
$$

where the inequality $(B2)$ is due to the relation (A.13) when $v = j$. This result contradicts the constraint $\widetilde{x}_j^* \leq \sum_{l=1}^n Q_{jl}$. Therefore, the location k cannot be the critical location in Scenario 2.

Part C. We prove that any location q ($q \neq i, j, k$) cannot be the critical location in Scenario 2. Suppose not, and the location q is the critical location in Scenario 2. Then we have $\widetilde{x}_q^* = \sum_{l=1}^n Q_{ql}$. According to Lemma 2, we have $X_q'(t) \geq \widetilde{x}_q^*,\ \forall t \geq 1$. Because we know $X_q'(t) \leq \sum_{l=1}^n Q_{ql} = \widetilde{x}_q^*,\ \forall t \geq 1$, we have that $X_q'(t) = \widetilde{x}_q^* = \sum_{l=1}^n Q_{ql},\ \forall t \geq 1$.

We next show that the following relation holds.

$$
X_v'(t) \geq x_v^*,\ \forall v \neq i, k, q,\ \forall t \geq 1 \tag{A.14a}
$$

$$
X_i'(t) \geq x_i^* + \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\},\ \forall t \geq 1 \tag{A.14b}
$$

$$
X_k'(t) \geq x_k^* + \min\left\{ \sum_m Q_{km} - x_k^*,\ \Phi_k \right\},\ \forall t \geq 1 \tag{A.14c}
$$

$$
X_q'(t) = \sum_{l=1}^{n} Q_{ql},\ \forall t \geq 1 \tag{A.14d}
$$

where

$$
\Phi_i = \left(\sum_m Q_{qm} - x_q^* \right) \cdot \alpha_{qi} + \min\left\{ \left(\sum_m Q_{qm} - x_q^* \right) \cdot \alpha_{qk}, \sum_m Q_{km} - x_k^* \right\} \cdot \alpha_{ki}
$$

$$
\Phi_k = \left(\sum_m Q_{qm} - x_q^* \right) \cdot \alpha_{qk} + \min\left\{ \left(\sum_m Q_{qm} - x_q^* \right) \cdot \alpha_{qi}, \sum_m Q_{im} - x_i^* \right\} \cdot \alpha_{ik}
$$

We already proven the last relation in (A.14). We just need to prove the first three relations. It is straightforward to see that the relation (A.14) holds for $t = 1$. Suppose the relation (A.14) holds for the time period t, we next show that it also holds for the time period $t+1$.

For the location i, we have

$$
\begin{aligned}
& X_i'(t+1) \\
=& \min\left\{ \sum_{m=1}^{n} Q_{im},\ UI_i'(t) + \sum_{m} Y_{mi}'(t) \right\} \\
\geq& \min\left\{ \sum_{m=1}^{n} Q_{im},\ Y_{ki}'(t) + Y_{qi}'(t) + \sum_{m\neq k,q} Y_{mi}'(t) \right\} \\
=& \min\left\{ \sum_{m=1}^{n} Q_{im},\ \sum_{m\neq k,q} X_m'(t)\cdot \frac{Q_{mi}}{\sum_{l=1}^{n} Q_{ml}} \right. \\
& \qquad \left. + X_k'(t)\cdot \frac{Q_{ki}}{\sum_{l=1}^{n} Q_{kl}} + X_q'(t)\cdot \frac{Q_{qi}}{\sum_{l=1}^{n} Q_{ql}} \right\} \\
\geq& \min\left\{ \sum_{m=1}^{n} Q_{im},\ \sum_{m\neq k,q} x_m^*\cdot \frac{Q_{mi}}{\sum_{l=1}^{n} Q_{ml}} \right. \\
& \qquad + \left(x_k^* + \min\left\{ \sum_{m} Q_{km} - x_k^*,\ \Phi_k \right\} \right)\cdot \frac{Q_{ki}}{\sum_{l=1}^{n} Q_{kl}} \\
& \qquad \left. + \left(x_q^* + \sum_{l=1}^{n} Q_{ql} - x_q^* \right)\cdot \frac{Q_{qi}}{\sum_{l=1}^{n} Q_{ql}} \right\} \\
=& \min\left\{ \sum_{m=1}^{n} Q_{im},\ \sum_{m} x_m^*\cdot \frac{Q_{mi}}{\sum_{l=1}^{n} Q_{ml}} + \alpha_{ki}\cdot \min\left\{ \sum_{m} Q_{km} - x_k^*,\ \Phi_k \right\} + \alpha_{qi}\cdot \left(\sum_{l=1}^{n} Q_{ql} - x_q^* \right) \right\} \\
=& \min\left\{ \sum_{m=1}^{n} Q_{im},\ x_i^* + \alpha_{ki}\cdot \min\left\{ \sum_{m} Q_{km} - x_k^*,\ \Phi_k \right\} + \alpha_{qi}\cdot \left(\sum_{l=1}^{n} Q_{ql} - x_q^* \right) \right. \\
=& x_i^* + \min\left\{ \sum_{m=1}^{n} Q_{im} - x_i^*,\ \alpha_{ki}\cdot \min\left\{ \sum_{m} Q_{km} - x_k^*,\ \Phi_k \right\} + \alpha_{qi}\cdot \left(\sum_{l=1}^{n} Q_{ql} - x_q^* \right) \right. \\
\overset{(C1)}{\geq}& x_i^* + \min\left\{ \sum_{m} Q_{im} - x_i^*,\ \Phi_i \right\},
\end{aligned}
$$

where the inequality $(C1)$ is due to the following relation

$$
\begin{aligned}
&\Phi_k - \left(\sum_m Q_{qm} - x_q^*\right)\cdot\alpha_{qk} = \min\left\{\left(\sum_m Q_{qm} - x_q^*\right)\cdot\alpha_{qi}, \sum_m Q_{im} - x_i^*\right\}\cdot\alpha_{ik} > 0 \\
\Longrightarrow\ &\min\left\{\sum_m Q_{km} - x_k^*,\ \Phi_k\right\} - \min\left\{\left(\sum_m Q_{qm} - x_q^*\right)\cdot\alpha_{qk}, \sum_m Q_{km} - x_k^*\right\} \geq 0 \\
\Longrightarrow\ &\alpha_{ki}\cdot\min\left\{\sum_m Q_{km} - x_k^*,\ \Phi_k\right\} - \min\left\{\left(\sum_m Q_{qm} - x_q^*\right)\cdot\alpha_{qk}, \sum_m Q_{km} - x_k^*\right\}\cdot\alpha_{ki} \geq 0 \\
\Longrightarrow\ &\alpha_{ki}\cdot\min\left\{\sum_m Q_{km} - x_k^*,\ \Phi_k\right\} + \alpha_{qi}\cdot\left(\sum_{l=1}^n Q_{ql} - x_q^*\right) \\
&\qquad - \left(\sum_m Q_{qm} - x_q^*\right)\cdot\alpha_{qi} - \min\left\{\left(\sum_m Q_{qm} - x_q^*\right)\cdot\alpha_{qk}, \sum_m Q_{km} - x_k^*\right\}\cdot\alpha_{ki} \geq 0 \\
\Longrightarrow\ &\alpha_{ki}\cdot\min\left\{\sum_m Q_{km} - x_k^*,\ \Phi_k\right\} + \alpha_{qi}\cdot\left(\sum_{l=1}^n Q_{ql} - x_q^*\right) - \Phi_i \geq 0
\end{aligned}
$$

For the location k, we have

$$
\begin{aligned}
&X_k'(t+1) \\
=&\min\left\{\sum_{m=1}^n Q_{km},\ UI_k'(t) + \sum_m Y_{mk}'(t)\right\} \\
\geq&\min\left\{\sum_{m=1}^n Q_{km},\ Y_{ik}'(t) + Y_{qk}'(t) + \sum_{m\neq i,q} Y_{mk}'(t)\right\} \\
=&\min\left\{\sum_{m=1}^n Q_{km},\ \sum_{m\neq i,q} X_m'(t)\cdot\frac{Q_{mk}}{\sum_{l=1}^n Q_{ml}}\right. \\
&\qquad\qquad \left. + X_i'(t)\cdot\frac{Q_{ik}}{\sum_{l=1}^n Q_{il}} + X_q'(t)\cdot\frac{Q_{qk}}{\sum_{l=1}^n Q_{ql}}\right\} \\
\geq&\min\left\{\sum_{m=1}^n Q_{km},\ \sum_{m\neq i,q} x_m^*\cdot\frac{Q_{mk}}{\sum_{l=1}^n Q_{ml}}\right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \left(x_i^* + \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} \right) \cdot \frac{Q_{ik}}{\sum_{l=1}^n Q_{il}} \\
&\quad + \left(x_q^* + \sum_{l=1}^n Q_{ql} - x_q^* \right) \cdot \frac{Q_{qk}}{\sum_{l=1}^n Q_{ql}} \Bigg\} \\
&= \min\left\{ \sum_{m=1}^n Q_{km},\ \sum_m x_m^* \cdot \frac{Q_{mk}}{\sum_{l=1}^n Q_{ml}} + \alpha_{ik} \cdot \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} + \alpha_{qk} \cdot \left(\sum_{l=1}^n Q_{ql} - x_q^* \right) \right\} \\
&= \min\left\{ \sum_{m=1}^n Q_{km},\ x_k^* + \alpha_{ik} \cdot \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} + \alpha_{qk} \cdot \left(\sum_{l=1}^n Q_{ql} - x_q^* \right) \right. \\
&= x_k^* + \min\left\{ \sum_{m=1}^n Q_{km} - x_k^*,\ \alpha_{ik} \cdot \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} + \alpha_{qk} \cdot \left(\sum_{l=1}^n Q_{ql} - x_q^* \right) \right. \\
&\overset{(C2)}{\geq} x_k^* + \min\left\{ \sum_m Q_{km} - x_k^*,\ \Phi_k \right\},
\end{aligned}
$$

where the inequality $(C2)$ is due to the following relation

$$
\begin{aligned}
&\Phi_i - \left(\sum_m Q_{qm} - x_q^* \right) \cdot \alpha_{qi} = \min\left\{ \left(\sum_m Q_{qm} - x_q^* \right) \cdot \alpha_{qk}, \sum_m Q_{km} - x_k^* \right\} \cdot \alpha_{ki} > 0 \\
\implies &\min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} - \min\left\{ \left(\sum_m Q_{qm} - x_q^* \right) \cdot \alpha_{qi}, \sum_m Q_{im} - x_i^* \right\} \geq 0 \\
\implies &\alpha_{ik} \cdot \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} - \min\left\{ \left(\sum_m Q_{qm} - x_q^* \right) \cdot \alpha_{qi}, \sum_m Q_{im} - x_i^* \right\} \cdot \alpha_{ik} \geq 0 \\
\implies &\alpha_{ik} \cdot \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} + \alpha_{qk} \cdot \left(\sum_{l=1}^n Q_{ql} - x_q^* \right) \\
&\qquad - \left(\sum_m Q_{qm} - x_q^* \right) \cdot \alpha_{qk} - \min\left\{ \left(\sum_m Q_{qm} - x_q^* \right) \cdot \alpha_{qi}, \sum_m Q_{im} - x_i^* \right\} \cdot \alpha_{ik} \geq 0 \\
\implies &\alpha_{ik} \cdot \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} + \alpha_{qk} \cdot \left(\sum_{l=1}^n Q_{ql} - x_q^* \right) - \Phi_k \geq 0
\end{aligned}
$$

For any location $v \neq i,k,q$, we have

$$
\begin{aligned}
& X'_v(t+1) \\
=& \min\left\{ \sum_{m=1}^{n} Q_{vm},\ UI'_v(t) + \sum_{m} Y'_{mv}(t) \right\} \\
\geq& \min\left\{ \sum_{m=1}^{n} Q_{vm},\ Y'_{iv}(t) + Y'_{kv}(t) + Y'_{qv}(t) + \sum_{m\neq i,k,q} Y'_{mv}(t) \right\} \\
=& \min\left\{ \sum_{m=1}^{n} Q_{vm},\ \sum_{m\neq i,k,q} X'_m(t) \cdot \frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}} \right. \\
& \qquad \left. + X'_i(t) \cdot \frac{Q_{iv}}{\tau + \sum_{l=1}^{n} Q_{il}} + X'_k(t) \cdot \frac{Q_{kv}}{\sum_{l=1}^{n} Q_{kl}} + X'_q(t) \cdot \frac{Q_{qv}}{\sum_{l=1}^{n} Q_{ql}} \right\} \\
\geq& \min\left\{ \sum_{m=1}^{n} Q_{vm},\ \sum_{m\neq i,k,q} x^*_m \cdot \frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}} \right. \\
& \qquad + \left(x^*_i + \min\left\{ \sum_{m} Q_{im} - x^*_i,\ \Phi_i \right\} \right) \cdot \frac{Q_{iv}}{\tau + \sum_{l=1}^{n} Q_{il}} \\
& \qquad + \left(x^*_k + \min\left\{ \sum_{m} Q_{km} - x^*_k,\ \Phi_k \right\} \right) \cdot \frac{Q_{kv}}{\sum_{l=1}^{n} Q_{kl}} \\
& \qquad \left. + \left(x^*_q + \sum_{l=1}^{n} Q_{ql} - x^*_q \right) \cdot \frac{Q_{qv}}{\sum_{l=1}^{n} Q_{ql}} \right\} \\
=& \min\left\{ \sum_{m=1}^{n} Q_{vm},\ \sum_{m} x^*_m \cdot \frac{Q_{mv}}{\sum_{l=1}^{n} Q_{ml}} + \alpha_{kv} \cdot \min\left\{ \sum_{m} Q_{km} - x^*_k,\ \Phi_k \right\} + \alpha_{qv} \cdot \left(\sum_{l=1}^{n} Q_{ql} - x^*_q \right) \right. \\
& \qquad \left. + x^*_i \cdot \left(\frac{Q_{iv}}{\tau + \sum_{l=1}^{n} Q_{il}} - \frac{Q_{iv}}{\sum_{l=1}^{n} Q_{il}} \right) + \frac{Q_{iv}}{\tau + \sum_{l=1}^{n} Q_{il}} \cdot \min\left\{ \sum_{m} Q_{im} - x^*_i,\ \Phi_i \right\} \right\} \\
=& \min\left\{ \sum_{m=1}^{n} Q_{vm},\ x^*_v + \alpha_{kv} \cdot \min\left\{ \sum_{m} Q_{km} - x^*_k,\ \Phi_k \right\} + \alpha_{qv} \cdot \left(\sum_{l=1}^{n} Q_{ql} - x^*_q \right) \right. \\
& \qquad \left. + x^*_i \cdot \left(\frac{Q_{iv}}{\tau + \sum_{l=1}^{n} Q_{il}} - \frac{Q_{iv}}{\sum_{l=1}^{n} Q_{il}} \right) + \frac{Q_{iv}}{\tau + \sum_{l=1}^{n} Q_{il}} \cdot \min\left\{ \sum_{m} Q_{im} - x^*_i,\ \Phi_i \right\} \right\}
\end{aligned}
$$

$$\overset{(C3)}{\geq} \min\left\{\sum_{m=1}^{n} Q_{vm},\ x_v^*\right\}$$

$$= x_v^*$$

where the inequality $(C3)$ is due to the following relation

$$\tau \leq \frac{\sum_{l=1}^{n} Q_{il}}{x_i^*} \cdot \min\left\{\sum_m Q_{im} - x_i^*,\ \Phi_i\right\} \tag{A.15a}$$

$$+ \frac{\sum_{l=1}^{n} Q_{il}}{x_i^*} \cdot \frac{\alpha_{kv}}{\alpha_{iv}} \cdot \min\left\{\sum_m Q_{km} - x_k^*,\ \Phi_k\right\} + \frac{\sum_{l=1}^{n} Q_{il}}{x_i^*} \cdot \frac{\alpha_{qv}}{\alpha_{iv}} \cdot \left(\sum_{l=1}^{n} Q_{ql} - x_q^*\right) \tag{A.15b}$$

$$\implies \frac{1}{\sum_{l=1}^{n} Q_{il}} \cdot x_i^* \cdot \alpha_{iv} \cdot \tau \leq \frac{Q_{iv}}{\sum_{l=1}^{n} Q_{il}} \cdot \min\left\{\sum_m Q_{im} - x_i^*,\ \Phi_i\right\} \tag{A.15c}$$

$$+ \alpha_{kv} \cdot \min\left\{\sum_m Q_{km} - x_k^*,\ \Phi_k\right\} + \alpha_{qv} \cdot \left(\sum_{l=1}^{n} Q_{ql} - x_q^*\right) \tag{A.15d}$$

$$\implies \alpha_{kv} \cdot \min\left\{\sum_m Q_{km} - x_k^*,\ \Phi_k\right\} + \alpha_{qv} \cdot \left(\sum_{l=1}^{n} Q_{ql} - x_q^*\right) \geq \tag{A.15e}$$

$$\frac{1}{\sum_{l=1}^{n} Q_{il}} \cdot \left(x_i^* \cdot \alpha_{iv} \cdot \tau - Q_{iv} \cdot \min\left\{\sum_m Q_{im} - x_i^*,\ \Phi_i\right\}\right) \tag{A.15f}$$

$$\implies \left[\sum_{l=1}^{n} Q_{il}\right] \cdot \left[\alpha_{kv} \cdot \min\left\{\sum_m Q_{km} - x_k^*,\ \Phi_k\right\} + \alpha_{qv} \cdot \left(\sum_{l=1}^{n} Q_{ql} - x_q^*\right)\right] \geq \tag{A.15g}$$

$$x_i^* \cdot \alpha_{iv} \cdot \tau - Q_{iv} \cdot \min\left\{\sum_m Q_{im} - x_i^*,\ \Phi_i\right\} \tag{A.15h}$$

$$\implies \left[\tau + \sum_{l=1}^{n} Q_{il}\right] \cdot \left[\alpha_{kv} \cdot \min\left\{\sum_m Q_{km} - x_k^*,\ \Phi_k\right\} + \alpha_{qv} \cdot \left(\sum_{l=1}^{n} Q_{ql} - x_q^*\right)\right] > \tag{A.15i}$$

$$x_i^* \cdot \alpha_{iv} \cdot \tau - Q_{iv} \cdot \min\left\{\sum_m Q_{im} - x_i^*,\ \Phi_i\right\} \tag{A.15j}$$

$$\implies \alpha_{kv} \cdot \min\left\{\sum_m Q_{km} - x_k^*,\ \Phi_k\right\} + \alpha_{qv} \cdot \left(\sum_{l=1}^{n} Q_{ql} - x_q^*\right) > \tag{A.15k}$$

$$\frac{1}{\tau+\sum_{l=1}^{n} Q_{il}} \cdot \left(x_i^* \cdot \alpha_{iv} \cdot \tau - Q_{iv} \cdot \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} \right) \quad \text{(A.15l)}$$

$$\implies \alpha_{kv} \cdot \min\left\{ \sum_m Q_{km} - x_k^*,\ \Phi_k \right\} + \alpha_{qv} \cdot \left(\sum_{l=1}^{n} Q_{ql} - x_q^* \right) > \quad \text{(A.15m)}$$

$$+ x_i^* \cdot \frac{\tau \cdot Q_{iv}}{(\sum_{l=1}^{n} Q_{il}) \cdot (\tau + \sum_{l=1}^{n} Q_{il})} - \frac{Q_{iv}}{\tau+\sum_{l=1}^{n} Q_{il}} \cdot \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} \quad \text{(A.15n)}$$

$$\implies \alpha_{kv} \cdot \min\left\{ \sum_m Q_{km} - x_k^*,\ \Phi_k \right\} + \alpha_{qv} \cdot \left(\sum_{l=1}^{n} Q_{ql} - x_q^* \right) > \quad \text{(A.15o)}$$

$$+ x_i^* \cdot \left(\frac{Q_{iv}}{\sum_{l=1}^{n} Q_{il}} - \frac{Q_{iv}}{\tau+\sum_{l=1}^{n} Q_{il}} \right) - \frac{Q_{iv}}{\tau+\sum_{l=1}^{n} Q_{il}} \cdot \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} \quad \text{(A.15p)}$$

Therefore, we have proven that the relation (A.14) holds for all locations in the network. Then we have

$$\begin{aligned}
\widetilde{x}_v^* &= \lim_{t\to\infty} X_v'(t) \geq x_v^*,\ \forall v \neq i,k,q, \\
\widetilde{x}_i^* &= \lim_{t\to\infty} X_i'(t) \geq x_i^* + \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\}, \\
\widetilde{x}_k^* &= \lim_{t\to\infty} X_k'(t) \geq x_k^* + \min\left\{ \sum_m Q_{km} - x_k^*,\ \Phi_k \right\}, \\
\widetilde{x}_q^* &= \sum_{l=1}^{n} Q_{ql}.
\end{aligned}$$

Further, we have

$$\begin{aligned}
\widetilde{x}_j^* - x_j^* &= \sum_{m\neq j} \widetilde{x}_m^* \cdot \alpha_{mj}' - x_j^* \\
&= \widetilde{x}_i^* \cdot \alpha_{ij}' + \widetilde{x}_k^* \cdot \alpha_{kj}' + \widetilde{x}_q^* \cdot \alpha_{qj}' + \sum_{m\neq i,k,q} \widetilde{x}_m^* \cdot \alpha_{mj}' - x_j^* \\
&\geq \widetilde{x}_i^* \cdot \alpha_{ij}' + \widetilde{x}_k^* \cdot \alpha_{kj}' + \widetilde{x}_q^* \cdot \alpha_{qj}' + \sum_{m\neq i,k,q} x_m^* \cdot \alpha_{mj}' - x_j^*
\end{aligned}$$

$$
\begin{aligned}
&= (\widetilde{x}_i^* - x_i^*) \cdot \alpha'_{ij} + (\widetilde{x}_k^* - x_k^*) \cdot \alpha'_{kj} + (\widetilde{x}_q^* - x_q^*) \cdot \alpha'_{qj} + \sum_m x_m^* \cdot \alpha'_{mj} - x_j^* \\
&= (\widetilde{x}_i^* - x_i^*) \cdot \alpha'_{ij} + (\widetilde{x}_k^* - x_k^*) \cdot \alpha'_{kj} + (\widetilde{x}_q^* - x_q^*) \cdot \alpha'_{qj} \\
&\geq \alpha'_{ij} \cdot \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} + \alpha'_{kj} \cdot \min\left\{ \sum_m Q_{km} - x_k^*,\ \Phi_k \right\} + \alpha'_{qj} \cdot \left(\sum_{l=1}^n Q_{ql} - x_q^* \right) \\
&= \alpha_{kj} \cdot \min\left\{ \sum_m Q_{km} - x_k^*,\ \Phi_k \right\} + \alpha_{qj} \cdot \left(\sum_{l=1}^n Q_{ql} - x_q^* \right) \\
&\qquad + \frac{Q_{ij}}{\tau + \sum_{l=1}^n Q_{il}} \cdot \min\left\{ \sum_m Q_{im} - x_i^*,\ \Phi_i \right\} \\
&\overset{(C4)}{>} x_i^* \cdot \left(\frac{Q_{ij}}{\sum_{l=1}^n Q_{il}} - \frac{Q_{ij}}{\tau + \sum_{l=1}^n Q_{il}} \right) \\
&> 0,
\end{aligned}
$$

where the inequality $(C4)$ is due to the relation (A.15) when $v = j$. So we have $\widetilde{x}_j^* > x_j^* = \sum_m Q_{jm}$, which contradicts the constraint $\widetilde{x}_j^* \leq \sum_m Q_{jm}$. Therefore, any location q ($q \neq i, j, k$) cannot be the critical location in Scenario 2.

From Part A, B and C, we have proven that all locations except the location j cannot be the critical location in Scenario 2. Therefore, only location j can be the critical location in Scenario 2. Hence proven. □

A.8 Proof of Proposition 8

Proof. Proof of Proposition 8 We first prove that $x_s \leq \sum_{k=1}^n d_k \cdot Q_{sk} \cdot \min\{r_{min}, 1\}$. From constraints (2.5b) and (2.5d), we have for any $j = 1, \cdots, n$, the following holds.

$$
x_j = x_s \cdot \frac{d_j \cdot Q_{sj}}{\sum_{k=1}^n d_k \cdot Q_{sk}} \leq d_j \cdot Q_{js},\ \forall j = 1, \cdots, n
$$

$$\implies x_s \leq \sum_{k=1}^{n} d_k \cdot Q_{sk} \cdot \frac{Q_{js}}{Q_{sj}},\ \forall j = 1, \cdots, n$$
$$\implies x_s \leq \sum_{k=1}^{n} d_k \cdot Q_{sk} \cdot r_{min},\ \forall j = 1, \cdots, n$$

Together with constraint (2.5e), we have that x_s cannot be more than $\sum_{k=1}^{n} d_k \cdot Q_{sk} \cdot \min\{r_{min}, 1\}$.

Further, we have $x_j = x_s \cdot \frac{d_j \cdot Q_{sj}}{\sum_{k=1}^{n} d_k \cdot Q_{sk}} \leq d_j \cdot Q_{sj} \cdot \min\{r_{min}, 1\}, \forall j = 1, \cdots, n$. Therefore, the maximum total network flow must be no more than $x_s + \sum_{j=1}^{n} d_j \cdot x_j = 2 \cdot \sum_{k=1}^{n} d_k \cdot Q_{sk} \cdot \min\{r_{min}, 1\}$.

Next, we show the upper bounds for x_js, x_s and the total network flow are achievable. The following solution achieves the upper bounds; its feasibility is easy to check.

$$x_j = d_j \cdot Q_{sj} \cdot \min\{r_{min}, 1\}, \forall j = 1, \cdots, n$$
$$x_s = \sum_{k=1}^{n} d_k \cdot Q_{sk} \cdot \min\{r_{min}, 1\}$$

Next, we prove the first result. Suppose the threshold strucrure is not true, then in the optimal solution we can find a location $j' \neq j_0$ such that $\frac{Q_{j's}}{Q_{sj'}} \geq r_0$ and $d_{j'} = 0$ hold. We now construct a new solution $\widetilde{\boldsymbol{d}}$ as follows: $\widetilde{d}_j = d_j, \forall j \neq j'$; $\widetilde{d}_j = 1, j = j'$. Next we prove that the total network flow for the new solution $\widetilde{\boldsymbol{d}}$ is higher than $\boldsymbol{d}$.

Suppose the optimal network flow for $\boldsymbol{d}$ is $\boldsymbol{x}$. We construct a flow solution $\widetilde{\boldsymbol{x}}$ for $\widetilde{\boldsymbol{d}}$ as follows: $\widetilde{x}_j = x_j, \forall j \neq j', \widetilde{x}_{j'} = Q_{sj'} \cdot \min\{r_{min}, 1\}$, and $\widetilde{x}_s = x_s + Q_{sj'} \cdot \min\{r_{min}, 1\}$. Next, we prove this solution is feasible for the subset selection solution $\widetilde{\boldsymbol{d}}$. We check the feasibility of each constraint in (2.5).

$$\text{(2.5b)}: \ \widetilde{x}_j - \widetilde{x}_s \cdot \frac{\widetilde{d}_j \cdot Q_{sj}}{\sum_{k=1}^{n} \widetilde{d}_k \cdot Q_{sk}}$$
$$= \widetilde{d}_j \cdot Q_{sj} \cdot \min\{r_{min}, 1\}$$

$$
\begin{aligned}
&- \sum_{k=1}^{n} \widetilde{d}_k \cdot Q_{sk} \cdot \min\{r_{min}, 1\} \cdot \frac{\widetilde{d}_j \cdot Q_{sj}}{\sum_{k=1}^{n} \widetilde{d}_k \cdot Q_{sk}} \\
=&0, \forall j = 1, \cdots, n \\
\text{(2.5c)}: \ &\widetilde{x}_s - \sum_{k=1}^{n} \widetilde{d}_k \cdot \widetilde{x}_k \\
=&x_s + Q_{sj'} \cdot \min\{r_{min}, 1\} - \sum_{k=1}^{n} d_k \cdot x_k - Q_{sj'} \cdot \min\{r_{min}, 1\} \\
=&0, \\
\text{(2.5d)}: \ &\widetilde{x}_j = x_j = d_j \cdot Q_{js} \leq \widetilde{d}_j \cdot Q_{js}, \forall j \neq j', \\
&\widetilde{x}_{j'} \leq \widetilde{d}_{j'} \cdot Q_{j's}, \\
\text{(2.5e)}: \ &\widetilde{x}_s = x_s + Q_{sj'} \cdot \min\{r_{min}, 1\} \\
&\leq \sum_{j \neq j'} d_j \cdot Q_{sj} + Q_{sj'} \cdot \min\{r_{min}, 1\} \\
&\leq \sum_{j=1}^{n} \widetilde{d}_j \cdot Q_{sj}.
\end{aligned}
$$

Therefore, by including the location j' in the service region, the total network flow will increase; this contradicts the optimality of $\boldsymbol{x}$ and $\boldsymbol{d}$. Hence proven. □

A.9 Proof of Proposition 9

Proof. Proof of Proposition 9

First we present a Lemma that is needed for our proof.

Lemma 3. *In the optimal solution of problem* (2.6)*,*

1. $(z_{sj} + \sum_k z_{kj}) \cdot (z_{js} + \sum_k z_{jk}) = 0, \forall j = 1, \cdots, n;$
2. $\sum_k z_{ks} \cdot \sum_k z_{sk} = 0;$

3. $\left(z_{js}+\sum_k z_{jk}\right)\cdot\left[\left(\beta_j+(n-1)\cdot\theta\right)\cdot(1-F(p_T))-x_j\right]=0,\forall j=1,\cdots,n;$

4. $\left(\sum_k z_{sk}\right)\cdot\left[\beta_s\cdot(1-F(p_T))-x_s\right]=0.$

Because the network operator should either reposition vehicles out from a location or into the location but should never do both to a same location, the first two results in Lemma 3 hold. To prove the third result, we discuss two cases. Case 1: if a location j does not have excessive vehicles, i.e., $\left(\beta_j+(n-1)\cdot\theta\right)\cdot(1-F(p_T))-x_j>0$, then the network operator should not reposition vehicles from it, i.e., $z_{js}+\sum_k z_{jk}=0$ must hold. Case 2: $\left(\beta_j+(n-1)\cdot\theta\right)\cdot(1-F(p_T))-x_j=0$, and the third result must hold. The proof for the fourth result is similar.

In the following, we prove Proposition 9. First, we prove there exists j_1 such that $\Delta_j>0,\forall j\leq j_1$. It suffices to prove that for $j,l=1,\cdots,n$, if $\Delta_j>0$ and $j>l$, then $\Delta_l>0$ holds. Assume $\Delta_j>0$, then we must have $\left(z_{js}+\sum_k z_{jk}\right)>0$. From Lemma 3, we know that $x_j=\left(\beta_j+(n-1)\cdot\theta\right)\cdot(1-F(p_T))$.

From the equation (2.6b), we have

$$x_j-x_l=x_l\cdot\frac{\theta}{\beta_l+(n-1)\cdot\theta}-x_j\cdot\frac{\theta}{\beta_j+(n-1)\cdot\theta}-\Delta_j+\Delta_l, \tag{A.16a}$$

$$\Longrightarrow\quad \Delta_l=-x_l\cdot\frac{\theta+\beta_l+(n-1)\cdot\theta}{\beta_l+(n-1)\cdot\theta}+x_j\cdot\frac{\theta+\beta_j+(n-1)\cdot\theta}{\beta_j+(n-1)\cdot\theta}+\Delta_j, \tag{A.16b}$$

$$=-x_l\cdot\frac{\theta+\beta_l+(n-1)\cdot\theta}{\beta_l+(n-1)\cdot\theta}+(\theta+\beta_j+(n-1)\cdot\theta)\cdot(1-F(p_T))+\Delta_j, \tag{A.16c}$$

$$\geq-(\theta+\beta_l+(n-1)\cdot\theta)\cdot(1-F(p_T))+(\theta+\beta_j+(n-1)\cdot\theta)\cdot(1-F(p_T))+\Delta_j, \tag{A.16d}$$

$$>0, \tag{A.16e}$$

Because $\Delta_l>0$, we have $\left(z_{ls}+\sum_k z_{lk}\right)>0$. From Lemma 3, we know that $x_l=\left(\beta_l+(n-1)\cdot\theta\right)\cdot(1-F(p_T))$. Therefore, we have proven that there exists j_1 such that $\Delta_j>0,\forall j\leq j_1$.

Next, we prove the existence of j_2, which then completes our proof.

First, we prove that if $\Delta_l < 0$, then $x_j = \left(\beta_j + (n-1)\cdot\theta\right)\cdot(1-F(p_T)), \forall j \geq l+1$. Suppose not, then we can find $j > l, \Delta_l < 0$ and $x_j < \left(\beta_j + (n-1)\cdot\theta\right)\cdot(1-F(p_T))$. Denote the associated solution as $p_T, \Delta_s, \Delta_1, \cdots, \Delta_n$ and $\boldsymbol{x}$. From the proof in the previous part, we know that Δ_j cannot be positive because otherwise Δ_l would also be positive. Therefore, $\Delta_j \leq 0$.

We have

$$\begin{aligned}
x_s =& \sum_{k=1}^{n} x_k \cdot \frac{\beta_k}{\beta_k + (n-1)\cdot\theta} - \Delta_s \\
=& \sum_{k\neq j} x_k \cdot \frac{\beta_k}{\beta_k + (n-1)\cdot\theta} + x_j \cdot \frac{\beta_j}{\beta_j + (n-1)\cdot\theta} - \Delta_s \\
<& \sum_{k\neq j} \beta_k + \beta_j - \Delta_s \\
=& \beta_s - \Delta_s
\end{aligned}$$

If $\Delta_s > 0$, then we have $x_s < \beta_s$, which violates Lemma 3. Therefore, we must have $\Delta_s \leq 0$. Next we argue that $x_1 = \left(\beta_1 + (n-1)\cdot\theta\right)\cdot(1-F(p_T))$ must hold. We discuss three cases. Case 1: $\Delta_s < 0$. In this case, there must be vehicles repositioned from spokes. As a result, we have $j_1 \geq 1$ and $\Delta_1 > 0$; then $x_1 = \left(\beta_1 + (n-1)\cdot\theta\right)\cdot(1-F(p_T))$ must hold according to Lemma 3. Case 2: $\Delta_s = 0$, and $\Delta_{t_1} < 0$ for some spoke t_1. In this case, there must be vehicles repositioned from spokes. As a result, we have $j_1 \geq 1$ and $\Delta_1 > 0$; then $x_1 = \left(\beta_1 + (n-1)\cdot\theta\right)\cdot(1-F(p_T))$ must hold according to Lemma 3. Case 3: $\Delta_s = 0$ and there is no reposition of vehicles in the entire network. In this case, because spoke 1 is the critical location, we must have $x_1 = \left(\beta_1 + (n-1)\cdot\theta\right)\cdot(1-F(p_T))$.

Because $\Delta_l < 0$, we then know that l cannot be 1. Next, we construct a new solution with strictly higher revenue than the above solution. We discuss two cases.

Case 1: $x_s < \beta_s \cdot (1-F(p_T))$. Keeping the price to be the same as p_T, we construct a new

solution in the following, which we prove has greater total network flow.

$$\begin{aligned}
\Delta'_s &= \Delta_s \\
\Delta'_l &= \Delta_l + \varepsilon_1 \\
\Delta'_j &= \Delta_j - \varepsilon_1 \\
\Delta'_j &= \Delta_j, \ \ \forall j \neq j, l.
\end{aligned}$$

where ε_1 satisfies

$$0 < \varepsilon_1 < \min\left\{ -\Delta_l,\ \left(\beta_j + (n-1)\cdot\theta\right)\cdot(1-F(p_T)) - x_j,\ \beta_s\cdot(1-F(p_T)) - x_s \right\} \tag{A.17a}$$

We next verify the following flow solution is the equilibrium flow with the above reposition solution.

$$\begin{aligned}
x'_s =& x_s + \varepsilon_1 \cdot \frac{\beta_j}{\theta + \beta_j + (n-1)\cdot\theta} - \varepsilon_1 \cdot \frac{\beta_l}{\theta + \beta_l + (n-1)\cdot\theta} \\
=& x_s - \varepsilon_1 \cdot \frac{n\cdot\theta}{\theta + \beta_j + (n-1)\cdot\theta} + \varepsilon_1 \cdot \frac{n\cdot\theta}{\theta + \beta_l + (n-1)\cdot\theta} \\
x'_l =& x_l - \varepsilon_1 \cdot \frac{\beta_l + (n-1)\cdot\theta}{\theta + \beta_l + (n-1)\cdot\theta} \\
x'_j =& x_j + \varepsilon_1 \cdot \frac{\beta_j + (n-1)\cdot\theta}{\theta + \beta_j + (n-1)\cdot\theta} \\
x'_j =& x_j, \ \ \forall j \neq j, l, s.
\end{aligned}$$

For the constraints (2.6b), for all $m = 1, \cdots, n$, if $m \neq j$ or l, we have

$$\frac{x'_s}{n} + \sum_{k=1, k\neq m}^{n} x'_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} - \Delta'_m$$

$$
\begin{aligned}
=&\frac{x_s}{n}-\varepsilon_1\cdot\frac{\theta}{\theta+\beta_j+(n-1)\cdot\theta}+\varepsilon_1\cdot\frac{\theta}{\theta+\beta_l+(n-1)\cdot\theta}\\
&+\sum_{k=1,k\neq m}^{n}x_k\cdot\frac{\theta}{\beta_k+(n-1)\cdot\theta}\\
&-\varepsilon_1\cdot\frac{\theta}{\theta+\beta_l+(n-1)\cdot\theta}+\varepsilon_1\cdot\frac{\theta}{\theta+\beta_j+(n-1)\cdot\theta}\\
&-\Delta_m\\
=&\frac{x_s}{n}+\sum_{k=1,k\neq m}^{n}x_k\cdot\frac{\theta}{\beta_k+(n-1)\cdot\theta}-\Delta_m\\
=&x'_j
\end{aligned}
$$

If $m=j$, we have

$$
\begin{aligned}
&\frac{x'_s}{n}+\sum_{k=1,k\neq m}^{n}x'_k\cdot\frac{\theta}{\beta_k+(n-1)\cdot\theta}-\Delta'_m\\
=&\frac{x'_s}{n}+\sum_{k=1,k\neq j}^{n}x'_k\cdot\frac{\theta}{\beta_k+(n-1)\cdot\theta}-\Delta'_j\\
=&\frac{x_s}{n}-\varepsilon_1\cdot\frac{\theta}{\theta+\beta_j+(n-1)\cdot\theta}+\varepsilon_1\cdot\frac{\theta}{\theta+\beta_l+(n-1)\cdot\theta}\\
&+\sum_{k=1,k\neq j}^{n}x_k\cdot\frac{\theta}{\beta_k+(n-1)\cdot\theta}\\
&-\varepsilon_1\cdot\frac{\theta}{\theta+\beta_l+(n-1)\cdot\theta}\\
&-\Delta_j+\varepsilon_1\\
=&\frac{x_s}{n}+\sum_{k=1,k\neq j}^{n}x_k\cdot\frac{\theta}{\beta_k+(n-1)\cdot\theta}-\Delta_j+\varepsilon_1\cdot\frac{\beta_j+(n-1)\cdot\theta}{\theta+\beta_j+(n-1)\cdot\theta}\\
=&x_j+\varepsilon_1\cdot\frac{\beta_j+(n-1)\cdot\theta}{\theta+\beta_j+(n-1)\cdot\theta}\\
=&x'_j
\end{aligned}
$$

If $m = l$, we have

$$
\begin{aligned}
&\frac{x'_s}{n} + \sum_{k=1,k\neq m}^{n} x'_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} - \Delta'_m \\
=&\frac{x'_s}{n} + \sum_{k=1,k\neq l}^{n} x'_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} - \Delta'_l \\
=&\frac{x_s}{n} - \varepsilon_1 \cdot \frac{\theta}{\theta + \beta_j + (n-1)\cdot\theta} + \varepsilon_1 \cdot \frac{\theta}{\theta + \beta_l + (n-1)\cdot\theta} \\
&+ \sum_{k=1,k\neq l}^{n} x_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} \\
&+ \varepsilon_1 \cdot \frac{\theta}{\theta + \beta_j + (n-1)\cdot\theta} \\
&- \Delta_l - \varepsilon_1 \\
=&\frac{x_s}{n} + \sum_{k=1,k\neq j}^{n} x_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} - \Delta_j - \varepsilon_1 \cdot \frac{\beta_l + (n-1)\cdot\theta}{\theta + \beta_l + (n-1)\cdot\theta} \\
=&x_l - \varepsilon_1 \cdot \frac{\beta_l + (n-1)\cdot\theta}{\theta + \beta_l + (n-1)\cdot\theta} \\
=&x'_l
\end{aligned}
$$

Therefore, we have proven the constraints (2.6b) hold for the new solution.

For the constraints (2.6c), we have

$$
\begin{aligned}
&\sum_{k=1}^{n} x'_k \cdot \frac{\beta_k}{\beta_k + (n-1)\cdot\theta} - \Delta'_s \\
=&\sum_{k=1}^{n} x_k \cdot \frac{\beta_k}{\beta_k + (n-1)\cdot\theta} - \varepsilon_1 \cdot \frac{\beta_l}{\theta + \beta_l + (n-1)\cdot\theta} + \varepsilon_1 \cdot \frac{\beta_j}{\theta + \beta_j + (n-1)\cdot\theta} - \Delta_s \\
=&x_s + \varepsilon_1 \cdot \frac{\beta_j}{\theta + \beta_j + (n-1)\cdot\theta} - \varepsilon_1 \cdot \frac{\beta_l}{\theta + \beta_l + (n-1)\cdot\theta} \\
=&x'_s
\end{aligned}
$$

Therefore, the constraint (2.6c) holds for the new solution.

It is easy to verify that the constraints (2.6d) and (2.6e) are satisfied because of the condition (A.17a). In the new solution, all the demands at spoke 1 are also satisfied because $x'_1 = x_1 = \big(\beta_1 + (n-1)\cdot\theta\big)\cdot(1-F(p_T))$.

The reposition costs of the old solution and the new solution are the same because the only difference between the two solutions is that ε_1 amount of repositioned vehicles to the spoke l are shifted to the spoke j. This also means that the new solution satisfies the constraint (2.6f).

The new solution has more network flows than the old solution because

$$\begin{aligned}
&\sum_{j=1}^{n} x'_j + x'_s - \sum_{j=1}^{n} x_j - x_s \\
=&\varepsilon_1 \cdot \frac{\beta_j}{\theta+\beta_j+(n-1)\cdot\theta} - \varepsilon_1 \cdot \frac{\beta_l}{\theta+\beta_l+(n-1)\cdot\theta} \\
&+ \varepsilon_1 \cdot \frac{\beta_j+(n-1)\cdot\theta}{\theta+\beta_j+(n-1)\cdot\theta} - \varepsilon_1 \cdot \frac{\beta_l+(n-1)\cdot\theta}{\theta+\beta_l+(n-1)\cdot\theta} \\
\geq& 0.
\end{aligned}$$

We have proven that the new solution has more revenue. Therefore, the old solution is not optimal, which contradicts our assumption.

Case 2: $x_s = \beta_s \cdot (1-F(p_T))$. In this case, we have

$$\begin{aligned}
\Delta_s &= \sum_{k=1}^{n} x_k \cdot \frac{\beta_k}{\beta_k+(n-1)\cdot\theta} - x_s \\
&< \big(\beta_j+(n-1)\cdot\theta\big)\cdot(1-F(p_T))\cdot\frac{\beta_k}{\beta_k+(n-1)\cdot\theta} + \sum_{k=1,k\neq j}^{n} x_k \cdot \frac{\beta_k}{\beta_k+(n-1)\cdot\theta} - x_s \\
&\leq \sum_{k=1}^{n} \beta_k \cdot (1-F(p_T)) - x_s
\end{aligned}$$

$$= \sum_{k=1}^{n} \beta_k \cdot (1 - F(p_T)) - \beta_s \cdot (1 - F(p_T))$$

$$= 0$$

Keeping the price to be the same as p_T, we construct a new solution in the following, which we prove has greater total network flow.

$$\Delta'_s = \Delta_s + \varepsilon_s$$

$$\Delta'_l = \Delta_l + \varepsilon_2$$

$$\Delta'_j = \Delta_j - \varepsilon_2 - \varepsilon_s$$

$$\Delta'_k = \Delta_k, \ \ \forall k \neq j, l.$$

where $\varepsilon_2 > 0, \varepsilon_s \geq 0$ and they satisfy

$$\varepsilon_s < -\Delta_s, \tag{A.18a}$$

$$\varepsilon_2 < -\Delta_l, \tag{A.18b}$$

$$\varepsilon_2 + \varepsilon_s < \big(\beta_j + (n-1) \cdot \theta\big) \cdot (1 - F(p_T)) - x_j, \tag{A.18c}$$

$$\varepsilon_2 \cdot \left(\frac{\beta_j}{\theta + \beta_j + (n-1) \cdot \theta} - \frac{\beta_l}{\theta + \beta_l + (n-1) \cdot \theta} \right) = \varepsilon_s \cdot \left(1 - \frac{\beta_j}{\theta + \beta_j + (n-1) \cdot \theta} \right) \tag{A.18d}$$

From the condition (A.18d), we have

$$\varepsilon_2 \cdot \left(- \frac{n \cdot \theta}{\theta + \beta_j + (n-1) \cdot \theta} + \frac{n \cdot \theta}{\theta + \beta_l + (n-1) \cdot \theta} \right) = \varepsilon_s \cdot \left(\frac{n \cdot \theta}{\theta + \beta_j + (n-1) \cdot \theta} \right) \tag{A.19a}$$

Further, we have

$$\varepsilon_2 \cdot \frac{\theta}{\theta+\beta_l+(n-1)\cdot\theta} = (\varepsilon_2+\varepsilon_s)\cdot\frac{\theta}{\theta+\beta_j+(n-1)\cdot\theta} \tag{A.20a}$$

We next verify the following flow solution is the equilibrium flow with the above reposition solution.

$$\begin{aligned}
x'_s =& x_s \\
x'_l =& x_l - \varepsilon_2 \cdot \frac{\beta_l+(n-1)\cdot\theta}{\theta+\beta_l+(n-1)\cdot\theta} \\
x'_j =& x_j + (\varepsilon_2+\varepsilon_s)\cdot\frac{\beta_j+(n-1)\cdot\theta}{\theta+\beta_j+(n-1)\cdot\theta} \\
x'_j =& x_j, \ \ \forall j \neq j,l,s.
\end{aligned}$$

For the constraints (2.6b), for all $m = 1,\cdots,n$, if $m \neq j$ or l, we have

$$\begin{aligned}
&\frac{x'_s}{n} + \sum_{k=1,k\neq m}^{n} x'_k \cdot \frac{\theta}{\beta_k+(n-1)\cdot\theta} - \Delta'_m \\
=&\frac{x_s}{n} + \sum_{k=1,k\neq m}^{n} x_k \cdot \frac{\theta}{\beta_k+(n-1)\cdot\theta} \\
&- \varepsilon_2 \cdot \frac{\theta}{\theta+\beta_l+(n-1)\cdot\theta} + (\varepsilon_2+\varepsilon_s)\cdot\frac{\theta}{\theta+\beta_j+(n-1)\cdot\theta} - \Delta_m \\
=&\frac{x_s}{n} + \sum_{k=1,k\neq m}^{n} x_k \cdot \frac{\theta}{\beta_k+(n-1)\cdot\theta} - \Delta_m \\
=&x'_j,
\end{aligned}$$

where in the second equality, we used the relation (A.20a).

If $m = j$, we have

$$\begin{aligned}
&\frac{x'_s}{n} + \sum_{k=1,k\neq m}^{n} x'_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} - \Delta'_m \\
=&\frac{x'_s}{n} + \sum_{k=1,k\neq j}^{n} x'_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} - \Delta'_j \\
=&\frac{x_s}{n} + \sum_{k=1,k\neq j}^{n} x_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} \\
&- \varepsilon_2 \cdot \frac{\theta}{\theta + \beta_l + (n-1)\cdot\theta} - \Delta_j + \varepsilon_2 + \varepsilon_s \\
=&\frac{x_s}{n} + \sum_{k=1,k\neq j}^{n} x_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} - \Delta_j \\
&- (\varepsilon_2 + \varepsilon_s) \cdot \frac{\theta}{\theta + \beta_j + (n-1)\cdot\theta} + \varepsilon_2 + \varepsilon_s \\
=&x_j + (\varepsilon_2 + \varepsilon_s) \cdot \frac{\beta_j + (n-1)\cdot\theta}{\theta + \beta_j + (n-1)\cdot\theta} \\
=&x'_j
\end{aligned}$$

If $m = l$, we have

$$\begin{aligned}
&\frac{x'_s}{n} + \sum_{k=1,k\neq m}^{n} x'_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} - \Delta'_m \\
=&\frac{x'_s}{n} + \sum_{k=1,k\neq l}^{n} x'_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} - \Delta'_l \\
=&\frac{x_s}{n} + \sum_{k=1,k\neq l}^{n} x_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} - \Delta_l \\
&+ (\varepsilon_2 + \varepsilon_s) \cdot \frac{\theta}{\theta + \beta_j + (n-1)\cdot\theta} - \varepsilon_2 \\
=&\frac{x_s}{n} + \sum_{k=1,k\neq l}^{n} x_k \cdot \frac{\theta}{\beta_k + (n-1)\cdot\theta} - \Delta_l
\end{aligned}$$

$$
\begin{aligned}
&+\varepsilon_2 \cdot \frac{\theta}{\theta+\beta_l+(n-1)\cdot\theta} - \varepsilon_2 \\
=&x_l - \varepsilon_2 \cdot \frac{\beta_l+(n-1)\cdot\theta}{\theta+\beta_l+(n-1)\cdot\theta} \\
=&x_l'
\end{aligned}
$$

Therefore, we have proven the constraints (2.6b) hold for the new solution.

For the constraints (2.6c), we have

$$
\begin{aligned}
&\sum_{k=1}^{n} x_k' \cdot \frac{\beta_k}{\beta_k+(n-1)\cdot\theta} - \Delta_s' \\
=&\sum_{k=1}^{n} x_k \cdot \frac{\beta_k}{\beta_k+(n-1)\cdot\theta} - \varepsilon_2 \cdot \frac{\beta_l}{\theta+\beta_l+(n-1)\cdot\theta} + (\varepsilon_2+\varepsilon_s)\cdot\frac{\beta_j}{\theta+\beta_j+(n-1)\cdot\theta} - \Delta_s - \varepsilon_s \\
=&x_s - \varepsilon_2 \cdot \left(\frac{\beta_l}{\theta+\beta_l+(n-1)\cdot\theta} - \frac{\beta_j}{\theta+\beta_j+(n-1)\cdot\theta}\right) - \varepsilon_s \cdot \frac{n\cdot\theta}{\theta+\beta_j+(n-1)\cdot\theta} \\
=&x_s + \varepsilon_s \cdot \left(\frac{n\cdot\theta}{\theta+\beta_j+(n-1)\cdot\theta}\right) - \varepsilon_s \cdot \frac{n\cdot\theta}{\theta+\beta_j+(n-1)\cdot\theta} \\
=&x_s \\
=&x_s'.
\end{aligned}
$$

Therefore, the constraint (2.6c) holds for the new solution.

It is easy to verify that the constraints (2.6d) and (2.6e) are satisfied because of the condition (A.18). In the new solution, all the demands at spoke 1 are also satisfied because $x_1' = x_1 = \big(\beta_1+(n-1)\cdot\theta\big)\cdot(1-F(p_T))$.

The reposition costs of the old solution and the new solution are the same because the only differences between the two solutions are that ε_2 amount of repositioned vehicles to the spoke l are shifted to the spoke j, and that ε_s amount of repositioned vehicles to the spoke s are shifted to the spoke j. This also means that the new solution satisfies the constraint (2.6f).

The new solution has more network flows than the old solution because

$$\begin{aligned}
&\sum_{j=1}^{n} x'_j + x'_s - \sum_{j=1}^{n} x_j - x_s \\
=&(\varepsilon_2 + \varepsilon_s) \cdot \frac{\beta_j + (n-1)\cdot\theta}{\theta + \beta_j + (n-1)\cdot\theta} - \varepsilon_2 \cdot \frac{\beta_l + (n-1)\cdot\theta}{\theta + \beta_l + (n-1)\cdot\theta} \\
\geq& 0.
\end{aligned}$$

We have proven that the new solution has more revenue. Therefore, the old solution is not optimal, which contradicts our assumption.

We have proven that if $\Delta_l < 0$, then $x_j = \big(\beta_j + (n-1)\cdot\theta\big)\cdot(1-F(p_T)), \forall j \geq l+1$. Then for $j \geq l+1$, we have

$$\begin{aligned}
\Delta_j =& x_l \cdot \frac{\theta + \beta_l + (n-1)\cdot\theta}{\beta_l + (n-1)\cdot\theta} - x_j \cdot \frac{\theta + \beta_j + (n-1)\cdot\theta}{\beta_j + (n-1)\cdot\theta} + \Delta_l \\
=& x_l \cdot \frac{\theta + \beta_l + (n-1)\cdot\theta}{\beta_l + (n-1)\cdot\theta} - \big(\theta + \beta_j + (n-1)\cdot\theta\big)\cdot(1-F(p_T)) + \Delta_l \\
\leq& \big(\theta + \beta_l + (n-1)\cdot\theta\big)\cdot(1-F(p_T)) - \big(\theta + \beta_j + (n-1)\cdot\theta\big)\cdot(1-F(p_T)) + \Delta_l \\
<& 0
\end{aligned}$$

Therefore, there exists j_2, such that $\Delta_j < 0, \forall j \geq j_2$. This completes our proof. □

Appendix B

Appendix for Chapter 3

B.1 Proof of Proposition 10

We first show the validity of (3.2a) when $n = 1$. Suppose that patient j's position in provider k's schedule is $m_j + 1$. Then $x^k_{j,m_j+1} = 1$ and $ST^k_j = t^k_{m_j+1}$, and by implication, $x^k_{j,m+1} = 0$ for $m = 0, 1, \ldots, m_j - 1, m_j + 1, \ldots, m_k - 1$. Three cases are possible for the constraints in (3.2a).

(1) $m = 0, 1, \ldots, m_j - 1$. Here, $x^k_{j,m+1} = 0$ so the first summation on the RHS of (3.2a) becomes 0 and the second becomes $T_{max} - LT^k_{m^k - m_j + m}$. As a consequence, (3.2a) reduces to

$$t^k_{m_j+1} = ST^k_j \leq t^k_{m+1} + T_{max} - LT^k_{m^k - m_j + m}$$

Note that there are $m_j - m$ patient encounters for provider k between t^k_{m+1} and $t^k_{m_j+1}$, and that the remaining $m^k - (m_j - m)$ encounters start during the time period $(t^k_{m+1} - 0) + (T_{max} - t^k_{m_j+1})$. By definition, $LT^k_{m^k - m_j + m}$ equals the sum of the $m^k - m_j + m$ smallest service times of provider k's patients, which means that it is a lower bound on the total time for any combination of $m^k - (m_j - m)$ encounters. Accordingly, $LT^k_{m^k - m_j + m} \leq (t^k_{m+1} - 0) + (T_{max} - t^k_{m_j+1})$, which validates the above inequality.

(2) $m = m_j$. Here, $x^k_{j,m+1} = 1$ so both the first and second summation on the RHS of (3.2a) become 0. As a consequence, (3.2a) reduces to the following inequality given that $t^k_{m_j+1} = t^k_{m+1}$.

$$t^k_{m_j+1} = ST^k_j \leq t^k_{m+1}$$

(3) $m = m_j + 1, \ldots, m^k - 1$. Here, $x^k_{j,m+1} = 0$ so the first summation on the RHS of (3.2a) becomes $LT^k_{m-m_j}$ and the second becomes 0. Thus, (3.2a) reduces to the following.

$$t^k_{m_j+1} = ST^k_j \leq t^k_{m+1} - LT^k_{m-m_j}$$

For provider k, there are $m - m_j$ patient encounters starting between $t^k_{m_j+1}$ and t^k_{m+1}. Again by definition, $LT^k_{m-m_j}$ equals the sum of the $m - m_j$ smallest service times of provider k's patients and is a lower bound on the total time for any $m - m_j$ encounters. Therefore, we have $LT^k_{m-m_j} \leq t^k_{m+1} - t^k_{m_j+1}$, which validates the above inequality.

Next we prove that constraints (3.2a) are actually stronger than their counterparts in (3.1i) . For (3.2a) and any value of m between 0 and $m^k - 1$, we have

$$ST^k_j \leq t^k_{m+1} + \sum_{m' \leq m-1} (-LT^k_{m-m'}) \cdot x^k_{j,m'+1} + \sum_{m' \geq m+1} (T_{max} - LT^k_{m^k-m'+m}) \cdot x^k_{j,m'+1} \tag{B.1a}$$

$$\leq t^k_{m+1} + \sum_{m' \leq m-1} T_{max} \cdot x^k_{j,m'+1} + \sum_{m' \geq m+1} T_{max} \cdot x^k_{j,m'+1} \tag{B.1b}$$

$$= t^k_{m \cdot n^k + 1} + \left(\sum_{m' \neq m} x^k_{j,m \cdot n^k + 1} \right) \cdot T_{max} \tag{B.1c}$$

$$= t^k_{m \cdot n^k + 1} + (1 - x^k_{j,m \cdot n^k + 1}) \cdot T_{max},$$

$$m = 0, \ldots, m^k - 1,\ n^k = 1,\ j \in J,\ \forall\, k \in K(j) \tag{B.1d}$$

The proofs for the remaining inequalities are identical.

□

B.2 Proof of Proposition 11

We only provide a proof for the case in which there are two nurse practitioners. The arguments for the general case are similar. For the case with $n^1 = 2$ nurse practitioners, we need to show that if patient j_1 starts no later than patient j_2, then $ST^1_{j_2} - ST^1_{j_1} \geq \left(\sum_{i \in A_{j_2} \setminus A_{j_1}} s^1_i - \max_{i \in A_{j_2} \setminus A_{j_1}} s^1_i \right)/2$.

Suppose patients $i_1, i_2, \ldots, i_m \in A_{j_2} \setminus \left(A_{j_1} \cup \{j_1\} \right)$. All patients whose visit with a nurse practitioner is no earlier than $ST^1_{j_1}$ and no later than $ST^1_{j_2}$ are $j_1, i_1, i_2, \ldots, i_m, j_2$. Of these patients, there is at most one whose encounter with a nurse practitioner ends no earlier than $ST^1_{j_2}$ besides patient j_2. In other words, all patients $j_1, i_1, i_2, \ldots, i_m$ start no earlier than $ST_{j^1_1}$, and at most one them finishes no earlier than $ST^1_{j_2}$. Suppose that patient i^*, finishes no earlier than $ST^1_{j_2}$. Then,

$$\begin{aligned} 2 \cdot (ST^1_{j_2} - ST^1_{j_1}) &\geq s^1_{j_1} + s^1_{i_1} + s^1_{i_2} + \cdots + s^1_{i_m} - s^1_{i^*} \\ &\geq s^1_{j_1} + s^1_{i_1} + s^1_{i_2} + \cdots + s^1_{i_m} - \max\{s^1_{j_1}, s^1_{i_1}, s^1_{i_2}, \ldots, s^1_{i_m}\} \end{aligned}$$

or

$$ST^1_{j_2} - ST^1_{j_1} \geq \left(\sum_{i \in A_{j_2} \setminus A_{j_1}} s^1_i - \max_{i \in A_{j_2} \setminus A_{j_1}} s^1_i \right)/2$$

□

B.3 Room constraints

Additional decision variables

η_{ij} 1 if $T^1_i \leq T^1_j$, which means patient i is placed in a room no later than patient j is placed in a room (the rooms for i and j can be different), 0 otherwise

η'_{ij} 1 if $T_i^2 \leq T_j^1$, which means patient i finishes using a room no later than patient j starts to use a room (the rooms for i and j can be different), 0 otherwise

ζ_{ij} 1 if patients i and j use the same room, and patient j follows (not necessarily immediately) patient i, 0 otherwise

δ_j^r 1 if patient j uses room r, 0 otherwise

B.3.1 Entering-checking method

This method assures that there is a room available for patient j when service starts with his first provider. For any other patient i who has previously entered the clinic, we already know his starting time T_i^1 and ending time T_i^2, so we already know the values of η_{ij} and η'_{ij}. This information allows us to determine the number of occupied rooms when the patient j sees his first provider. To ensure that a room is available, this number must be less than the total number of rooms, R.

Proposition 19. *A necessary and sufficient condition that arriving patient j can be placed in a room is that* $\sum_{i\neq j, i\in J}\left(\eta_{ij}-\eta'_{ij}\right) \leq R-1, \forall j \in J$.

Proof. For any patient j whose visit starts with his first provider at time T_j^1, we need to show that the above inequality is satisfied if a room is available. That is, we need to determine how many of the R rooms are occupied. Now, for any other patient i, the three cases shown in Figure B.1 need to be considered.

(1) $T_i^1 < T_j^1$, $T_i^2 \leq T_j^1$. In this case, we have $\eta_{ij} = 1$ and $\eta'_{ij} = 1$, so the room used by patient i is available for patient j.

(2) $T_i^1 \leq T_j^1$, $T_i^2 > T_j^1$. In this case, we have $\eta_{ij} = 1$ and $\eta'_{ij} = 0$, indicating that patient i is still in his room so it is not available for patient j.

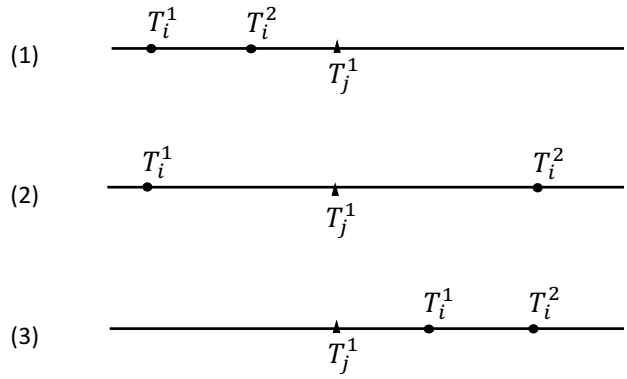


Figure B.1: An example for entering-checking method

(3) $T_i^1 > T_j^1$, $T_i^2 > T_j^1$. In this case, we have $\eta_{ij} = 0$ and $\eta'_{ij} = 0$, implying that patient i has not yet been placed in a room, so whichever room he is eventually assigned is immaterial to a room being available for patient j . Of course, the time in clinic for patients i and j may overlap, which implies that they cannot use the same room. This will be assured when a check is made for patient i to determine if a room is available, but it is not a concern when patient j is being assigned a room.

From these cases, we see that when patient j's encounter with his first provider begins, if $\eta_{ij} - \eta'_{ij} = 1$, then patient i is using a room; if $\eta_{ij} - \eta'_{ij} = 0$, then patient i is not using a room. Accordingly, when patient j enters the clinic at time T_j^1, the total number of rooms that are being used is $\sum_{i \neq j, i \in J} \left(\eta_{ij} - \eta'_{ij} \right)$. If patient j can be placed in a room, then the total number of rooms that are being used must be no more than $R-1$. In contrast, if the total number of rooms that are being used is R, then patient j cannot be placed in a room. Therefore, $\sum_{i \neq j, i \in J} \left(\eta_{ij} - \eta'_{ij} \right) \leq R-1$ is a necessary and sufficient condition that a room is available for patient j. □

Based on Proposition 19, we have the following constraints for the room requirement.

$$T_i^1 \geq T_j^1 - \eta_{ij} T_{max}, \qquad \forall\, i \neq j \in J \tag{B.2a}$$

$$T_j^1 \geq T_i^1 - (1-\eta_{ij}) \cdot T_{max}, \qquad \forall\, i \neq j \in J \tag{B.2b}$$

$$T_i^2 \geq T_j^1 - \eta'_{ij} T_{max}, \qquad \forall\, i \neq j \in J \tag{B.2c}$$

$$T_j^1 \geq T_i^2 - (1-\eta'_{ij}) \cdot T_{max}, \qquad \forall\, i \neq j \in J \tag{B.2d}$$

$$\sum_{i \neq j, i \in J} \left(\eta_{ij} - \eta'_{ij}\right) \leq R-1, \qquad \forall\, j \in J \tag{B.2e}$$

$$\eta_{ij} + \eta_{ji} \geq 1, \qquad \forall\, i \neq j \in J \tag{B.2f}$$

$$\eta_{ij} \geq \eta'_{ij}, \qquad \forall\, i \neq j \in J \tag{B.2g}$$

$$\sum_{m=1}^{m^k} m \cdot x_{jm}^k \geq \sum_{m=1}^{m^k} m \cdot x_{im}^k + 1 - m^k \cdot (1-\eta'_{ij}), \;\; \forall\, i, j \in J(k), \;\; \forall\, k \in \{k : n^k = 1\} \tag{B.2h}$$

$$\eta_{ij}, \eta'_{ij} \in \{0,1\}, \qquad \forall\, i, j \in J \tag{B.2i}$$

Constraints (B.2a) and (B.2b) ensure that $\eta_{ij} = 1$ when patient i is placed in a room no later than patient j, and 0 otherwise. Constraints (B.2c) and (B.2d) ensure $\eta'_{ij} = 1$ if patient i finishes using her room before patient j is placed in a room, and 0 otherwise. Constraints (B.2e) guarantee that the total number of rooms being used when patient j is placed in a room is no greater than $R-1$. Constraints (B.2f) specify that either patient i starts no later than j, or patient j starts no later than i. This is needed for the case in which patients i and j are placed in different rooms at the same time. Without (B.2f), η_{ij}, η'_{ij}, η_{ji} and η'_{ji} will all be 0 when rooming occurs simultaneously for the two patients.

Constraints (B.2g) are useful cuts which impose the restriction that if patient i finishes earlier than patient j starts, then patient i must also start earlier than patient j. Constraints (B.2h) are also useful cuts, which state that if patient i finishes earlier than patient j starts, then for any provider who is the only provider of her type, patient i's position index should be smaller than patient j's

position index. The difference must be at least 1. Constraints (B.2i) define the variables as binary.

B.3.2 Not-immediate-successor method

We begin by assigning each patient to a room. For any two patients who are assigned to a same room, we use binary variables to ensure that they don't overlap in time. That is, if two patients are assigned to the same room, then the starting time of the successor (not necessarily immediate successor) can be no earlier than the ending time of all his predecessors.

$$\sum_{r=1}^{R} \delta_j^r = 1, \qquad \forall\ j \in J \tag{B.3a}$$

$$\zeta_{ij} + \zeta_{ji} \geq \delta_i^r + \delta_j^r - 1, \qquad \forall\ i \neq j \in J,\ \ r = 1, \ldots, R \tag{B.3b}$$

$$T_j^1 \geq T_i^2 - (1 - \zeta_{ij}) \cdot T_{max}, \qquad \forall\ i \neq j \in J \tag{B.3c}$$

$$\sum_{m=1}^{m^k} m \cdot x_{jm}^k \geq \sum_{m=1}^{m^k} m \cdot x_{im}^k + 1 + \sum_{m \in J(k), m \neq i, m \neq j} (\zeta_{im} + \zeta_{mj} - 1) - (1 - \zeta_{ij}) \cdot m^k,$$
$$\forall\ i \neq j \in J(k),\ \forall\ k \in \{k : n^k = 1\} \tag{B.3d}$$

$$\zeta_{ij} \in \{0, 1\}, \qquad \forall\ i, j \in J \tag{B.3e}$$

Constraints (B.3a) ensure that each patient has a room. Constraints (B.3b) specify that two patients who are assigned to the same room must use the room in sequence. Constraints (B.3c) enforce the requirement that the starting time of patient j cannot be earlier than the ending time of all his predecessors i who are assigned the same room. Constraints (B.3d) and (B.3e) parallel (3.3d) and (3.3e) .

Appendix C

Appendix for Chapter 4

C.1 Proof of Proposition 12.

The proof is by contradiction. First, assume that the statement of the proposition is not true. Then in the optimal solution to $\beta\left(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\right)$, there must exist at least one k such that $U_{(k)} < Q_k^{(1-\varepsilon_k)}$. Let $k' = \arg\max_k \{k : U_{(k)} < Q_k^{(1-\varepsilon_k)}\}$ and denote the corresponding random variable to be $Z_{j(k')}$. Specifically, $j(k')$ is a mapping from order statistics index k' to the index j of random variable Z_j. We must have that $F_{j(k')}(Z_{j(k')}) = U_{(k')} < Q_{k'}^{(1-\varepsilon_{k'})}$.

Let $F_{j(k')}(Z_{j(k')} + \delta) = Q_{k'}^{(1-\varepsilon_{k'})}$. Such δ exists because $0 \leq Q_{k'}^{(1-\varepsilon_{k'})} \leq 1$; and $\delta > 0$ because $F_{j(k')}$ is non-decreasing. We construct a new solution to $\beta\left(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\right)$ by only modifying $Z_{j(k')}$ to be $Z_{j(k')} + \delta$. The new solution to $\beta\left(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\right)$ becomes $Z_1, \cdots, Z_{j(k')-1}, Z_{j(k')} + \delta, Z_{j(k')+1}, \cdots, Z_J$.

In the old solution to $\beta\left(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\right)$, the order statistics of $F_j(Z_j)$s are $U_{(1)}, \cdots, U_{(k'-1)}, U_{(k')}, U_{(k'+1)}, \cdots, U_{(J)}$. Note that in the new solution to $\beta\left(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\right)$, only $F_{j(k')}(Z_{j(k')}) = U_{(k')}$ increases to $F_{j(k')}(Z_{j(k')} + \delta)$, so for the new solution the order statistics of new $F_j(Z_j)$s become $U_{(1)}, \cdots, U_{(k'-1)}, F_{j(k')}(Z_{j(k')} + \delta), U_{(k'+1)}, \cdots, U_{(J)}$. Next, we show they are ordered from the smallest to the largest. The reasons are as follows: (1) because $U_{(k)} \leq U_{(k')} < F_{j(k')}(Z_{j(k')} + \delta), \forall k \leq k' - 1$, the first $k' - 1$ order statistics of the new $F_j(Z_j)$s are still $U_{(1)}, \cdots, U_{(k'-1)}$; (2) because $F_{j(k')}(Z_{j(k')} + \delta) = Q_{k'}^{(1-\varepsilon_{k'})} \leq U_{(k)}, \forall k \geq k' + 1$, we must have that the last $J - k'$ order statistics remain unchanged as $U_{(k'+1)}, \cdots, U_{(J)}$; (3) consequently, $F_{j(k')}(Z_{j(k')} + \delta)$ becomes the k'th order statistic of the new $F_j(Z_j)$s. Therefore,

we have proven that in the new solution to $\beta\big(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\big)$, the order statistics of $F_j(Z_j)$s are $U_{(1)}, \cdots, U_{(k'-1)}, F_{j(k')}(Z_{j(k')} + \delta), U_{(k'+1)}, \cdots, U_{(J)}$, and they satisfy the constraint in $\mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})$. This shows that the new solution $Z_1, \cdots, Z_{j(k')-1}, Z_{j(k')} + \delta, Z_{j(k')+1}, \cdots, Z_J$ is feasible to $\beta\big(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\big)$.

Because the new solution to $\beta\big(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\big)$ increases $\beta\big(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\big)$ by $\hat{a}_{j(k')}|x_{j(k')}| \cdot \delta \geq 0$, it means that the original solution to $\beta\big(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\big)$ was not a maximizer of $\beta\big(\mathbf{x}, \mathscr{U}'(\boldsymbol{\epsilon}', \boldsymbol{\epsilon})\big)$, and we must increase $Z_{j(k')}$ to be $Z_{j(k')} + \delta$. Using this same logic, we continue to modify our solution if there exist a k such that $U_{(k)} < Q_k^{(1-\varepsilon_k)}$, and eventually we must have $U_{(k)} = Q_k^{(1-\varepsilon_k)}, \forall k$. □

C.2 Proof of Proposition 13.

Without loss of generality, assume $k_1 < k_2$. As we stated in Section 4.2.1, we have $Q_{k_1}^{(1-\varepsilon_{k_1})} \leq Q_{k_2}^{(1-\varepsilon_{k_2})}$. Further due to $Q_{k_1}^{(1-\varepsilon_{k_1})} \neq Q_{k_2}^{(1-\varepsilon_{k_2})}$, we have $Q_{k_1}^{(1-\varepsilon_{k_1})} < Q_{k_2}^{(1-\varepsilon_{k_2})}$. Next we choose an arbitrary $\mathbf{Z} = (Z_1, \cdots, Z_j) \in \mathscr{U}^{OS}(\boldsymbol{\epsilon})$, such that there exist j_1, j_2 ($j_1 \neq j_2$) that satisfy $F_{j_1}(Z_{j_1}) = U_{j_1} = Q_{k_1}^{(1-\varepsilon_{k_1})}$ and $F_{j_2}(Z_{j_2}) = U_{j_2} = Q_{k_2}^{(1-\varepsilon_{k_1})}$.

Since $0 \leq Q_{k_1}^{(1-\varepsilon_{k_1})} \leq 1$ and $0 \leq Q_{k_2}^{(1-\varepsilon_{k_2})} \leq 1$, we can find W_1 and W_2, such that $F_{j_1}(W_1) = Q_{k_2}^{(1-\varepsilon_{k_2})}$ and $F_{j_2}(W_2) = Q_{k_1}^{(1-\varepsilon_{k_1})}$. Then we construct $\mathbf{Z}'$ by replacing Z_{j_1} and Z_{j_2} in $\mathbf{Z}$ with W_1 and W_2, respectively. So we have $Z'_{j_1} = W_1, Z'_{j_2} = W_2$ and $Z'_j = Z_j$, for $j \neq j_1, j_2$. Since $F_{j_1}(Z'_{j_1}) = F_{j_2}(Z_{j_2})$, $F_{j_2}(Z'_{j_2}) = F_{j_1}(Z_{j_1})$ and $F_j(Z'_j) = F_j(Z_j)$, for $j \neq j_1, j_2$, we have $\{F_1(Z'_1), \cdots, F_J(Z'_J)\} = \{F_1(Z_1), \cdots, F_J(Z_J)\}$. This shows that $\mathbf{Z}' \in \mathscr{U}^{OS}(\boldsymbol{\epsilon})$.

Because $Q_{k_1}^{(1-\varepsilon_{k_1})} < Q_{k_2}^{(1-\varepsilon_{k_2})}$, we have $Z_{j_1} < W_1$ and $Z_{j_2} > W_2$. Then for any $\lambda \in (0,1)$, we must have $F_{j_1}\big((1-\lambda)Z_{j_1} + \lambda W_1\big)\big) > Q_{k_1}^{(1-\varepsilon_{k_1})}$ and $F_{j_2}\big((1-\lambda)Z_{j_2} + \lambda W_2)\big) > Q_{k_1}^{(1-\varepsilon_{k_1})}$.

For any $\lambda \in (0,1)$, denote $\mathbf{Z}^\lambda = (1-\lambda)\mathbf{Z} + \lambda\mathbf{Z}'$. We next show $\mathbf{Z}^\lambda \notin \mathscr{U}^{OS}(\boldsymbol{\epsilon})$. Since $Z_j^\lambda = (1-\lambda)Z_j + \lambda Z'_j = Z_j$, for $j \neq j_1, j_2$, we have $\{F_j(Z_j^\lambda) : \forall j \neq j_1, j_2\} = \{F_j(Z_j) : \forall j \neq j_1, j_2\} = \left\{Q_k^{(1-\varepsilon_k)} : \forall k \neq k_1, k_2\right\}$. Therefore, the k_1-th order statistic of $\{F_j(Z_j^\lambda) : \forall j = 1, \cdots, J\} =$

$\min\left\{Q_{k_1+1}^{(1-\varepsilon/k_1+1)}, F_{j_1}\left((1-\lambda)Z_{j_1}+\lambda W_1\right), F_{j_2}\left((1-\lambda)Z_{j_2}+\lambda W_2\right)\right\}$ is strictly greater than $Q_{k_1}^{(1-\varepsilon_{k_1})}$. This proves that $\mathbf{Z}^\lambda \notin \mathscr{U}^{OS}(\boldsymbol{\epsilon})$, and our proof is completed. □

C.3 Proof of Proposition 14.

We prove the equivalence of problem (4.4) and $\beta\left(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\right)$ by showing that the optimal solution to each problem is feasible to the other.

First we prove the optimal solution to problem (4.4) is feasible to $\beta\left(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\right)$. Note that the problem (4.4) is essentially a linear relaxation of the maximum weight assignment problem. It is well-known that there is always an optimal solution with all the η variables taking integer values. In such optimal solution to problem (4.4), if $\eta_{jk} = 1$, we let $Z_j = \sum_{k\in J} q_{jk}\eta_{jk} = q_{jk}$. According to the definition of q_{jk}, we then have $F_j(Z_j)$=$F_j(q_{jk}) \leq Q_k^{(1-\varepsilon_k)}, \forall j,k \in J$. Therefore, for any $1 \leq m \leq J$, we must have the mth smallest element in the set $\{F_j(Z_j), \forall j \in J\}$ should be no greater than the the mth smallest element in the set $\{Q_k^{(1-\varepsilon_k)}, \forall k \in J\}$, which is essentially $U_{(m)} \leq Q_m^{(1-\varepsilon_m)}, \forall m \in J$. This proves that the optimal solution to problem (4.4) is indeed feasible to the problem $\beta\left(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\right)$.

Next we prove the optimal solution to $\beta\left(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\right)$ is feasible to problem (4.4). Assume in the optimal solution to $\beta\left(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{\epsilon})\right)$, the order statistics of $F_j(Z_j)$s are $F_{j_1}(Z_{j_1}), F_{j_2}(Z_{j_2}), \cdots, F_{j_J}(Z_{j_J})$, where $j_1, j_2, \cdots, j_J$ are a permutation of the set $\{1, 2, \cdots, J\}$. We then have $F_{j_k}(Z_{j_k}) = Q_k^{(1-\varepsilon_k)}, \forall k \in J$, i.e., $Z_{j_k} = q_{j_k,k}$. Such solution is feasible to problem (4.4) because we can construct the equivalent solution to problem (4.4) as follows:

$$\eta_{j_k,m} = \begin{cases} 1, & k = m, \\ 0, & k \neq m, \end{cases} \qquad \forall k, m \in J. \tag{C.1}$$

Hence proved. □

C.4 Proof of Proposition 15.

Similar to Bertsimas and Sim (2004), we apply the strong duality to reformulate Model (4.5). For fixed $\mathbf{x}$, we first take dual of the maximizing problem in constraints (4.5b), and we get:

$$\min \sum_{j\in J_i}(\theta_{ij}+\phi_{ij})+\sum_{j\in J_i}\sum_{k\in J_i}\zeta_{ijk} \tag{C.2a}$$

$$\text{s.t. } \theta_{ij}+\phi_{ik}+\zeta_{ijk}\geq \hat{a}_{ij}|x_j|q_{ijk}, \forall j,k\in J_i, \forall i \tag{C.2b}$$

$$\zeta_{ijk}\geq 0, \forall j,k\in J_i, \forall i \tag{C.2c}$$

Because the maximizing problem in constraints (4.5b) is feasible and bounded, we must have that the formulation (C.2) is also feasible and bounded due to strong duality. And their optimal objective values are equal. Substituting formulation (C.2) into Model (4.5), we can get the linear programming formulation (4.6). Hence proved. □

C.5 Proof: the equivalence of the RO models with the budget uncertainty set and the order statistic uncertainty set.

We prove that the RO model with the budget uncertainty set with budget τ (the Problem (4) in Bertsimas and Sim 2004) is equivalent to the RO model with the order statistic uncertainty set with properly chosen q_{jk} values.

The Problem (4) in Bertsimas and Sim (2004) is essentially the following problem. We assume that there is only one constraint that has the budget uncertainty set, and so we have removed the i

index.

$$\max_{\mathbf{x}} \ \mathbf{c}'\mathbf{x} \tag{C.3a}$$

$$\text{s.t.} \ \sum_j a_j x_j + \beta\left(\mathbf{x}, \mathscr{U}^B(\tau)\right) \leq b, \tag{C.3b}$$

$$\underline{\mathbf{x}} \leq \mathbf{x} \leq \overline{\mathbf{x}}, \tag{C.3c}$$

where

$$\beta\left(\mathbf{x}, \mathscr{U}^B(\tau)\right) = \max_{\{S\cup\{t\}|S\subset J, |S|=\lfloor\tau\rfloor, t\in J\setminus S\}} \left\{ \sum_{j\in S} \hat{a}_j \cdot |x_j| + (\tau - \lfloor\tau\rfloor)\cdot \hat{a}_t \cdot |x_t| \right\}. \tag{C.4a}$$

Note that we use $\underline{\mathbf{x}}$ and $\overline{\mathbf{x}}$ instead of $\mathbf{l}$ and $\mathbf{u}$ for lower and upper bounds, and τ for the budget instead of Γ. We need to prove that the problem (C.3) is equivalent to the following RO model with the order statistic uncertainty set.

$$\max_{\mathbf{x}} \ \mathbf{c}'\mathbf{x} \tag{C.5a}$$

$$\text{s.t.} \ \sum_j a_j x_j + \beta\left(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{q})\right) \leq b, \tag{C.5b}$$

$$\underline{\mathbf{x}} \leq \mathbf{x} \leq \overline{\mathbf{x}}, \tag{C.5c}$$

where

$$\beta\left(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{q})\right) = \max_{\eta} \ \sum_{j\in J} \hat{a}_j |x_j| \cdot \left(\sum_{k\in J} q_{jk}\eta_{jk} \right) \tag{C.6a}$$

$$\text{s.t.} \ \sum_k \eta_{jk} = 1, \forall j \in J \tag{C.6b}$$

$$\sum_j \eta_{jk} = 1, \forall k \in J \tag{C.6c}$$

$$0 \leq \eta_{jk} \leq 1, \forall j,k \in J. \tag{C.6d}$$

and $\boldsymbol{q}$ satisfies $q_{jk} = 0$, if $1 \leq k \leq J - \lfloor \tau \rfloor - 1, \forall j \in J$; $q_{jk} = \tau - \lfloor \tau \rfloor$, if $k = J - \lfloor \tau \rfloor, \forall j \in J$; $q_{jk} = 1$, if $J - \lfloor \tau \rfloor + 1 \leq k \leq J, \forall j \in J$.

We just need to prove that $\beta\big(\mathbf{x}, \mathscr{U}^B(\tau)\big) = \beta\big(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{q})\big)$.

The problem (C.6) is the linear relaxation of the maximum weight assignment problem, which is known to have an integer optimal solution. For every $j \in J$, there exists a unique $k \in J$ such that $\eta_{jk} = 1$. If $\eta_{jk} = 1$, then $\hat{a}_j|x_j|$ is paired to q_{jk}. Therefore, if $1 \leq k \leq J - \lfloor \tau \rfloor - 1$, then $\hat{a}_j|x_j|$ is paired with 0; if $k = J - \lfloor \tau \rfloor$, then $\hat{a}_j|x_j|$ is paired with $\tau - \lfloor \tau \rfloor$; if $J - \lfloor \tau \rfloor + 1 \leq k \leq J$, then $\hat{a}_j|x_j|$ is paired with 1. So for all $\hat{a}_1|x_1|, \hat{a}_2|x_2|, \cdots, \hat{a}_J|x_J|$, we know that $\lfloor \tau \rfloor$ of them will be paired with 1, and one of them will be paired with $\tau - \lfloor \tau \rfloor$, and the rest will be paired with 0. Therefore, the problem (C.6) is essentially the same as the problem (C.4), and so we have $\beta\big(\mathbf{x}, \mathscr{U}^B(\tau)\big) = \beta\big(\mathbf{x}, \mathscr{U}^{OS}(\boldsymbol{q})\big)$. Hence proven.

C.6 Proof of Proposition 16.

The constraints in $\mathscr{U}^D$ are equivalent to the following:

$$-\Gamma \cdot m^{1/\alpha} \leq \sum_{j \in S} Z_j \leq \Gamma \cdot m^{1/\alpha}, \forall S \subseteq J, |S| = m, m = 1, \cdots, J, \tag{C.7a}$$

$$\text{or} \quad -\Gamma \cdot m^{1/\alpha} \leq \min_{S \subseteq J, |S| = m} \sum_{j \in S} Z_j, \max_{S \subseteq J, |S| = m} \sum_{j \in S} Z_j \leq \Gamma \cdot m^{1/\alpha}, \forall m = 1, \cdots, J. \tag{C.7b}$$

Because the kth order statistic of Z_js is denoted as $Z_{(k)}$, we have the following:

$$\min_{S\subseteq J,|S|=m}\sum_{j\in S}Z_j=\sum_{k=1}^{m}Z_{(k)},\ \max_{S\subseteq J,|S|=m}\sum_{j\in S}Z_j=\sum_{k=J+1-m}^{J}Z_{(k)} \tag{C.8a}$$

Then the demand uncertainty set $\mathscr{U}^D$ can be rewritten in terms of the order statistics of Z_js as follows:

$$\mathscr{U}_{os}^D=\left\{\mathbf{Z}\in\mathbb{R}^J:-\Gamma j^{1/\alpha}\leq\sum_{k=1}^{j}Z_{(k)},\sum_{k=J+1-j}^{J}Z_{(k)}\leq\Gamma j^{1/\alpha},\forall j\in J\right\}.$$

Since the demand uncertainty set is equivalent to $\mathscr{U}_{os}^D$, we just need to prove that $\mathbf{Z}^*$ maximizes $\beta\left(\mathbf{x},\mathscr{U}_{os}^D\right)$. We first check $\mathbf{Z}^*$ is feasible to $\mathscr{U}_{os}^D$, and then prove it is optimal to the problem $\beta\left(\mathbf{x},\mathscr{U}_{os}^D\right)$.

To prove the feasibility of $\mathbf{Z}^*$, there are two steps: (1) we need to prove that $(J+1-k)^{1/\alpha}-(J-k)^{1/\alpha}$ is increasing in k. That is $[(J-k)^{1/\alpha}-(J-k-1)^{1/\alpha}]-[(J+1-k)^{1/\alpha}-(J-k)^{1/\alpha}]\geq 0,\forall 1\leq k\leq J-1$, or equivalently $2\cdot(J-k)^{1/\alpha}-(J-k-1)^{1/\alpha}-(J+1-k)^{1/\alpha}\geq 0$, which is evidenced by the fact that the function $x^{1/\alpha}$ is concave for $\alpha\geq 1$. (2) We check that $\mathbf{Z}^*$ satisfies all the constraints in $\mathscr{U}_{os}^D$. Note that $Z_{(k)}^*\geq 0,\forall k\in J$, so $-\Gamma j^{1/\alpha}\leq\sum_{k=1}^{j}Z_{(k)}^*,\forall j\in J$, is satisfied; for other constraints, $\sum_{k=J+1-j}^{J}Z_{(k)}^*=\sum_{k=J+1-j}^{J}\Gamma(J+1-k)^{1/\alpha}-\Gamma(J-k)^{1/\alpha}=\Gamma j^{1/\alpha}$. Therefore, $\mathbf{Z}^*$ is feasible to $\mathscr{U}_{os}^D$.

We next prove $\mathbf{Z}^*$ achieves optimal solution to the problem $\beta\left(\mathbf{x},\mathscr{U}_{os}^D\right)$. Denote the order statistics of $\hat{a}_1|x_1|,\hat{a}_2|x_2|,\cdots,\hat{a}_J|x_J|$ as $[\hat{a}|x|]_{(1)},[\hat{a}|x|]_{(2)},\cdots,[\hat{a}|x|]_{(J)}$, i.e., $[\hat{a}|x|]_{(k)}$ is the kth smallest among all $\hat{a}_j|x_j|$s. Because of the rearrangement inequality Cvetkovski (2012, Theorem 6.1), the corresponding objective value for any feasible $\mathbf{Z}$ in $\mathscr{U}_{os}^D$ should be no greater than $[\hat{a}|x|]_{(1)}\cdot Z_{(1)}+[\hat{a}|x|]_{(2)}\cdot Z_{(2)}+\cdots+[\hat{a}|x|]_{(J)}\cdot Z_{(J)}$. Then the difference of the optimal objective value of our $\mathbf{Z}^*$ and

that of any feasible $\mathbf{Z}$ in $\mathscr{U}_{os}^{D}$ should be no less than:

$$[\hat{a}|x|]_{(1)} \cdot (Z^*_{(1)} - Z_{(1)}) + [\hat{a}|x|]_{(2)} \cdot (Z^*_{(2)} - Z_{(2)}) + \cdots + [\hat{a}|x|]_{(J)} \cdot (Z^*_{(J)} - Z_{(J)}) \tag{C.9a}$$

Since $\mathbf{Z}$ satisfies $\sum_{k=J+1-j}^{J} Z_{(k)} \leq \Gamma j^{1/\alpha} = \sum_{k=J+1-j}^{J} Z^*_{(k)}, \forall j \in J$, we then have $\sum_{k=j}^{J}(Z^*_{(k)} - Z_{(k)}) \geq 0, \forall j \in J$, i.e., $Z^*_{(j)} - Z_{(j)} \geq -\sum_{k=j+1}^{J}(Z^*_{(k)} - Z_{(k)}), \forall 1 \leq j \leq J-1$. We apply these constraints to (C.9a) one at a time repeatedly, and we can get:

$$\begin{aligned}
\text{(C.9a)} \geq & -[\hat{a}|x|]_{(1)} \cdot \sum_{k=2}^{J}(Z^*_{(k)} - Z_{(k)}) + [\hat{a}|x|]_{(2)} \cdot (Z^*_{(2)} - Z_{(2)}) + \cdots + [\hat{a}|x|]_{(J)} \cdot (Z^*_{(J)} - Z_{(J)}) \\
= & ([\hat{a}|x|]_{(2)} - [\hat{a}|x|]_{(1)}) \cdot (Z^*_{(2)} - Z_{(2)}) + ([\hat{a}|x|]_{(3)} - [\hat{a}|x|]_{(1)}) \cdot (Z^*_{(3)} - Z_{(3)}) + \cdots \\
& + ([\hat{a}|x|]_{(J)} - [\hat{a}|x|]_{(1)}) \cdot (Z^*_{(J)} - Z_{(J)}) \\
\geq & -([\hat{a}|x|]_{(2)} - [\hat{a}|x|]_{(1)}) \cdot \sum_{k=3}^{J}(Z^*_{(k)} - Z_{(k)}) + ([\hat{a}|x|]_{(3)} - [\hat{a}|x|]_{(1)}) \cdot (Z^*_{(3)} - Z_{(3)}) + \cdots \\
& + ([\hat{a}|x|]_{(J)} - [\hat{a}|x|]_{(1)}) \cdot (Z^*_{(J)} - Z_{(J)}) \\
= & ([\hat{a}|x|]_{(3)} - [\hat{a}|x|]_{(2)}) \cdot (Z^*_{(3)} - Z_{(3)}) + ([\hat{a}|x|]_{(4)} - [\hat{a}|x|]_{(2)}) \cdot (Z^*_{(4)} - Z_{(4)}) + \cdots \\
& + ([\hat{a}|x|]_{(J)} - [\hat{a}|x|]_{(2)}) \cdot (Z^*_{(J)} - Z_{(J)}) \\
& \cdots \\
\geq & ([\hat{a}|x|]_{(J)} - [\hat{a}|x|]_{(J-1)}) \cdot (Z^*_{(J)} - Z_{(J)}) \\
\geq & 0
\end{aligned}$$

The last inequality holds because $Z_{(J)} \leq \Gamma = Z^*_{(J)}$, and $[\hat{a}|x|]_{(J)} - [\hat{a}|x|]_{(J-1)} \geq 0$ holds by definition. This concludes the proof. □

C.7 Proof of Corollary 2.

Similar to the problem (4.4), the problem (4.7) is also a relaxed maximum weight assignment problem. Therefore, there always exists an optimal solution where all the η variables take integer values. As a result, problem (4.7) is equivalent to $\max_{j_1,\cdots,j_J} \sum_{k\in J} \hat{a}_k|x_k| \cdot \Gamma\big({j_k}^{1/\alpha} - (j_k-1)^{1/\alpha}\big)$, where ${j_k}^{1/\alpha} - (j_k-1)^{1/\alpha}$ is assigned to $\hat{a}_k|x_k|$, and $\{j_1, j_2, \cdots, j_J\}$ is a permutation of the set $J = \{1,2,\cdots,J\}$. Due to the rearrangement inequality, the optimal solution to problem (4.7) must be $\sum_{k\in J} [\hat{a}|x|]_{(k)} \cdot \Gamma\big((J+1-k)^{1/\alpha} - (J-k)^{1/\alpha}\big)$, which is exactly the optimal solution to the problem $\beta\left(\mathbf{x}, \mathscr{U}^D\right)$ as we derived in the proof for Proposition 16. □

C.8 Proof of Proposition 18.

We denote $p(\boldsymbol{\epsilon}) = J!\det[\Delta]$. For any $A_1, A_2, \cdots, A_J$ that are independently and symmetrically distributed in $[a_j - \hat{a}_j, a_j + \hat{a}_j]$, we need to prove $\Pr\Big(\sum_{j\in J} A_j \cdot x_j \le b\Big) \ge \frac{1}{2} + \frac{1}{2} \cdot p(\boldsymbol{\epsilon})$. Denote $\rho_j = (A_j - a_j)/\hat{a}_j \in [-1,1]$ and recall $Z_j = |\rho_j|$, we then have

$$\Pr\left(\sum_j A_j x_j^* \le b\right) \tag{C.10a}$$

$$= \Pr\left(\sum_j a_j x_j^* + \sum_j \hat{a}_j |x_j^*| \cdot \rho_j \le b\right) \tag{C.10b}$$

$$= \Pr\left(\sum_j \hat{a}_j |x_j^*| \cdot \rho_j \le b - \sum_j a_j x_j^*\right) \tag{C.10c}$$

$$\ge \Pr\left(\sum_j \hat{a}_j |x_j^*| \cdot \rho_j \le \beta^*\right) \tag{C.10d}$$

Because ρ_j is symmetrically distributed in $[-1,1]$, we must have

$$\Pr\big(\sum_j \hat{a}_j|x_j^*|\cdot\rho_j < -\beta^*\big) = \Pr\big(\sum_j \hat{a}_j|x_j^*|\cdot\rho_j > \beta^*\big)$$

Further, we have

$$\Pr\left(\sum_j \hat{a}_j|x_j^*|\cdot\rho_j \le \beta^*\right) \tag{C.11a}$$

$$=\Pr\left(\sum_j \hat{a}_j|x_j^*|\cdot\rho_j < -\beta^*\right) + \Pr\left(\left|\sum_j \hat{a}_j|x_j^*|\cdot\rho_j\right| \le \beta^*\right) \tag{C.11b}$$

$$=\frac{1}{2}+\frac{1}{2}\cdot\Pr\left(\left|\sum_j \hat{a}_j|x_j^*|\cdot\rho_j\right| \le \beta^*\right) \tag{C.11c}$$

$$\ge\frac{1}{2}+\frac{1}{2}\cdot\Pr\left(\sum_j \hat{a}_j|x_j^*|\cdot|\rho_j| \le \beta^*\right) \tag{C.11d}$$

$$=\frac{1}{2}+\frac{1}{2}\cdot\Pr\left(\sum_j \hat{a}_j|x_j^*|\cdot Z_j \le \beta^*\right) \tag{C.11e}$$

$$=\frac{1}{2}+\frac{1}{2}\cdot\sum_{\left\{\{j_1,j_2,\cdots,j_J\}:\, F_{j_1}(Z_{j_1})\le F_{j_2}(Z_{j_2})\le\cdots\le F_{j_J}(Z_{j_J})\right\}} \frac{1}{J!}\Pr\left(\sum_{j\in\{j_1,j_2,\cdots,j_J\}} \hat{a}_j\cdot|x_j|\cdot Z_j \le \beta^*\right) \tag{C.11f}$$

where $\{j_1,j_2,\cdots,j_J\}$ is a permutation of the set $J=\{1,2,\cdots,J\}$. The last equality holds because the $F_j(Z_j)$ follows Unif$(0,1)$ distribution, and for any permutation $\{j_1,j_2,\cdots,j_J\}$, the probability of $F_{j_1}(Z_{j_1}) \le F_{j_2}(Z_{j_2}) \le \cdots \le F_{j_J}(Z_{j_J})$ is always equal to $\frac{1}{J!}$. We study each term in the sum in (C.11f).

$$\Pr\left(\sum_{j\in\{j_1,j_2,\cdots,j_J\}} \hat{a}_j\cdot|x_j|\cdot Z_j \le \beta^*\right)\Bigg|_{\left\{F_{j_1}(Z_{j_1})\le\cdots\le F_{j_J}(Z_{j_J})\right\}} \tag{C.12a}$$

$$\ge \max_{\left\{\chi:\, \sum_{j\in\{j_1,j_2,\cdots,j_J\}}\hat{a}_j\cdot|x_j|\cdot\chi_j\le\beta^*\right\}} \Pr\big(F_{j_k}(Z_{j_k}) \le F_{j_k}(\chi_{j_k}), \forall k\in J\big)\Big|_{\left\{F_{j_1}(Z_{j_1})\le\cdots\le F_{j_J}(Z_{j_J})\right\}} \tag{C.12b}$$

$$\geq \Pr\left(F_{j_k}(Z_{j_k}) \leq F_{j_k}(q_{j_k,k}), \forall k \in J\right)\Big|_{\left\{F_{j_1}(Z_{j_1}) \leq \cdots \leq F_{j_J}(Z_{j_J})\right\}} \tag{C.12c}$$

$$= \Pr\left(F_{j_k}(Z_{j_k}) \leq Q_k^{1-\varepsilon_k}), \forall k \in J\right)\Big|_{\left\{F_{j_1}(Z_{j_1}) \leq \cdots \leq F_{j_J}(Z_{j_J})\right\}} \tag{C.12d}$$

$$= \Pr\left(\cup_{k=1}^{J} \left\{U_{(k)} \leq Q_k^{1-\varepsilon_k}\right\}\right)\Big|_{U_j \sim \text{Unif}(0,1), \forall j \in J} \tag{C.12e}$$

$$= J! \det[\Delta] \tag{C.12f}$$

$$= p(\boldsymbol{\epsilon}) \tag{C.12g}$$

where the first inequality is obtained by applying Theorem 3.1 in Embrechts et al. (2003), and the second inequality holds because $\chi_{j_k} = q_{j_k,k}, \forall k$ is a feasible solution for χ defined by $\left\{\chi : \sum_{j \in \{j_1, j_2, \cdots, j_J\}} \hat{a}_j \cdot |x_j| \cdot \chi_j \leq \beta^*\right\}$. The second equality holds because $F_j(Z_j)$'s are independent uniform random variables. Then we have

$$\Pr\left(\sum_j A_j x_j^* \leq b\right) \geq \text{RHS of (C.11f)} \tag{C.13a}$$

$$\geq \frac{1}{2} + \frac{1}{2} \cdot p(\boldsymbol{\epsilon}) \cdot \frac{1}{J!} \cdot J! \tag{C.13b}$$

$$= \frac{1}{2} + \frac{1}{2} \cdot p(\boldsymbol{\epsilon}) \tag{C.13c}$$

□

Bibliography

Ahmadi-Javid, Amir, Zahra Jalali, Kenneth J Klassen. 2017. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research* **258**(1) 3–34.

Azadeh, Ali, Milad Baghersad, Mehdi Hosseinabadi Farahani, Mansour Zarrin. 2015. Semi-online patient scheduling in pathology laboratories. *Artificial Intelligence in Medicine* **64**(3) 217–226.

Bailey, Norman TJ. 1952. A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society. Series B (Methodological)* **14**(2) 185–199.

Bandi, Chaithanya, Dimitris Bertsimas. 2014. Robust option pricing. *European Journal of Operational Research* **239**(3) 842–853.

Bandi, Chaithanya, Dimitris Bertsimas, Nataly Youssef. 2015. Robust queueing theory. *Operations Research* **63**(3) 676–700.

Bandi, Chaithanya, Diwakar Gupta. 2019. Operating-room staffing and scheduling. *Manufacturing & Service Operations Management* to appear.

Banerjee, Siddhartha, Daniel Freund, Thodoris Lykouris. 2016. Pricing and optimization in shared vehicle systems: An approximation framework. *arXiv preprint arXiv:1608.06819* .

Bard, Jonathan F, David P Morton, Yong Min Wang. 2007. Workforce planning at usps mail processing and distribution centers using stochastic optimization. *Annals of Operations Research* **155**(1) 51.

Bard, Jonathan F, Zhichao Shu, Douglas J Morrice, Dongyang Wang, Ramin Poursani, Luci Leykum. 2016. Improving patient flow at a family health clinic. *Health Care Management Science* **19**(2) 170–191.

Ben-Tal, Aharon, Laurent El Ghaoui, Arkadi Nemirovski. 2009. *Robust optimization*, vol. 28. Princeton University Press.

Ben-Tal, Aharon, Arkadi Nemirovski. 1998. Robust convex optimization. *Mathematics of Operations Research* **23**(4) 769–805.

Ben-Tal, Aharon, Arkadi Nemirovski. 2000. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming* **88**(3) 411–424.

Berg, Bjorn P, Brian T Denton, S Ayca Erdogan, Thomas Rohleder, Todd Huschka. 2014. Optimal booking and scheduling in outpatient procedure centers. *Computers & Operations Research* **50** 24–37.

Bertsimas, Dimitris, Hoda Bidkhori. 2015. On the performance of affine policies for two-stage adaptive optimization: a geometric perspective. *Mathematical Programming* **153**(2) 577–594.

Bertsimas, Dimitris, David B Brown. 2009. Constructing uncertainty sets for robust linear optimization. *Operations Research* **57**(6) 1483–1495.

Bertsimas, Dimitris, David B Brown, Constantine Caramanis. 2011a. Theory and applications of robust optimization. *SIAM Review* **53**(3) 464–501.

Bertsimas, Dimitris, Vineet Goyal, Xu Andy Sun. 2011b. A geometric characterization of the power of finite adaptability in multistage stochastic and adaptive optimization. *Mathematics of Operations Research* **36**(1) 24–54.

Bertsimas, Dimitris, Vishal Gupta, Nathan Kallus. 2018. Data-driven robust optimization. *Mathematical Programming* **167**(2) 235–292.

Bertsimas, Dimitris, Melvyn Sim. 2004. The price of robustness. *Operations Research* **52**(1) 35–53.

Bhat, Prashanth B, Viktor K Prasanna, Cauligi S Raghavendra. 2000. Block-cyclic redistribution over heterogeneous networks. *Cluster Computing* **3**(1) 25–34.

Birge, John R, Francois Louveaux. 2011. *Introduction to stochastic programming*. Springer Science & Business Media.

Bosch, Peter M Vanden, Dennis C Dietz. 2000. Minimizing expected waiting in a medical appointment system. *IIE Transactions* **32**(9) 841–848.

Castro, Elkin, Sanja Petrovic. 2012. Combined mathematical programming and heuristics for a radiotherapy pre-treatment scheduling problem. *Journal of Scheduling* **15**(3) 333–346.

Cayirli, Tugba, Emre Veral. 2003. Outpatient scheduling in health care: A review of literature. *Production and Operations Management* **12**(4) 519–549.

Cayirli, Tugba, Emre Veral, Harry Rosen. 2006. Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science* **9**(1) 47–58.

Chakraborty, Santanu, Kumar Muthuraman, Mark Lawley. 2010. Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions* **42**(5) 354–366.

Chen, Rachel R, Lawrence W Robinson. 2014. Sequencing and scheduling appointments with potential call-in patients. *Production and Operations Management* **23**(9) 1522–1538.

CMS. 2018. `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.html`. Accessed 18 July 2018.

Cruz, Fábio, Anand Subramanian, Bruno P Bruck, Manuel Iori. 2017. A heuristic algorithm for a single vehicle static bike sharing rebalancing problem. *Computers & Operations Research* **79** 19–33.

Cvetkovski, Zdravko. 2012. *Inequalities: theorems, techniques and selected problems*. Springer Science & Business Media.

Delage, Erick, Yinyu Ye. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* **58**(3) 595–612.

Denton, Brian T, Andrew J Miller, Hari J Balasubramanian, Todd R Huschka. 2010. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research* **58**(4-part-1) 802–816.

Dielman, Terry, Cynthia Lowry, Roger Pfaffenberger. 1994. A comparison of quantile estimators. *Communications in Statistics-Simulation and Computation* **23**(2) 355–371.

Dobson, Gregory, Tolga Tezcan, Vera Tilson. 2013. Optimal workflow decisions for investigators in systems with interruptions. *Management Science* **59**(5) 1125–1141.

El Ghaoui, Laurent, Francois Oustry, Hervé Lebret. 1998. Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization* **9**(1) 33–52.

Elhenawy, Mohammed, Hesham Rakha. 2017. A heuristic for rebalancing bike sharing systems based on a deferred acceptance algorithm. *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. IEEE, 188–193.

Embrechts, Paul, Andrea Höing, Alessandro Juri. 2003. Using copulae to bound the value-at-risk for functions of dependent risks. *Finance and Stochastics* **7**(2) 145–167.

Engell, Sebastian, Andreas Märkert, Guido Sand, Rüdiger Schultz. 2004. Aggregated scheduling of a multiproduct batch plant by two-stage stochastic integer programming. *Optimization and Engineering* **5**(3) 335–359.

Erdogan, S Ayca, Brian Denton. 2013. Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS Journal on Computing* **25**(1) 116–132.

French, Kenneth R. 2019. Data library. `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. Accessed 18 March 2019.

Gao, Rui, Xi Chen, Anton J Kleywegt. 2017. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050* .

George, David K, Cathy H Xia. 2011. Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European journal of operational research* **211**(1) 198–207.

Glasserman, Paul, Philip Heidelberger, Perwez Shahabuddin. 2000. Variance reduction techniques for estimating value-at-risk. *Management Science* **46**(10) 1349–1364.

Gupta, Diwakar, Brian Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* **40**(9) 800–819.

Gupta, Diwakar, Lei Wang. 2008. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research* **56**(3) 576–592.

Gut, Allan. 2009. *An intermediate course in probability*. Springer-Verlag New York.

Hanna, Aura. 2010. Patient-centred care. *Ontario Medical Review* **1** 27.

He, Long, Ho-Yin Mak, Ying Rong, Zuo-Jun Max Shen. 2017. Service region design for urban electric vehicle sharing systems. *Manufacturing & Service Operations Management* **19**(2) 309–327.

Hong, Liang, Yu Zheng, Duncan Yung, Jingbo Shang, Lei Zou. 2015. Detecting urban black holes based on human mobility data. *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 1–10.

Jin, Li, Zhuonan Feng, Ling Feng. 2016. A context-aware collaborative filtering approach for urban black holes detection. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 2137–2142.

Kazemian, Pooyan, Mustafa Y Sir, Mark P Van Oyen, Jenna K Lovely, David W Larson, Kalyan S Pasupathy. 2017. Coordinating clinic and surgery appointments to meet access service levels for elective surgery. *Journal of Biomedical Informatics* **66** 105–115.

Keswani, Aakash, Karl M Koenig, Kevin J Bozic. 2016. Value-based healthcare: Part 1–designing and implementing integrated practice units for the management of musculoskeletal disease. *Clinical Orthopaedics and Related Research* **474**(10) 2100–2103.

Klabjan, Diego, David Simchi-Levi, Miao Song. 2013. Robust stochastic lot-sizing by means of histograms. *Production and Operations Management* **22**(3) 691–710.

Kleywegt, Anton J, Alexander Shapiro, Tito Homem-de Mello. 2002. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* **12**(2) 479–502.

Kong, Qingxia, Chung-Yee Lee, Chung-Piaw Teo, Zhichao Zheng. 2013. Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research* **61**(3) 711–726.

Lahiri, Atanu, Abraham Seidmann. 2012. Information hang-overs in healthcare service systems. *Manufacturing & Service Operations Management* **14**(4) 634–653.

Leeftink, A G, I A Bikker, I M H Vliegen, R J Boucherie. 2018. Multi-disciplinary planning in health care: A review. *Health Systems* doi:10.1080/20476965.2018.1436909.

Liaw, Ching-Fang. 2000. A hybrid genetic algorithm for the open shop scheduling problem. *European Journal of Operational Research* **124**(1) 28–42.

Lu, Mengshi, Zhihao Chen, Siqian Shen. 2018. Optimizing the profitability and quality of service in carshare systems under demand uncertainty. *Manufacturing & Service Operations Management* **20**(2) 162–180.

Mancilla, Camilo, Robert Storer. 2012. A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions* **44**(8) 655–670.

Mannino, Carlo, Eivind J Nilssen, Tomas Eric Nordlander. 2012. A pattern based, robust approach to cyclic master surgery scheduling. *Journal of Scheduling* **15**(5) 553–563.

Miller, Clair E, Albert W Tucker, Richard A Zemlin. 1960. Integer programming formulation of traveling salesman problems. *Journal of the ACM* **7**(4) 326–329.

NACTO. 2017. `https://nacto.org/bike-share-statistics-2017/`. Accessed 26 August 2020.

Noori-Darvish, Samane, Iraj Mahdavi, Nezam Mahdavi-Amiri. 2012. A bi-objective possibilistic programming model for open shop scheduling problems with sequence-dependent setup times, fuzzy processing times, and fuzzy due dates. *Applied Soft Computing* **12**(4) 1399–1416.

Oh, Hyun-Jung, Ana Muriel, Hari Balasubramanian, Katherine Atkinson, Thomas Ptaszkiewicz. 2013. Guidelines for scheduling in primary care under different patient types and stochastic nurse and provider service times. *IIE Transactions on Healthcare Systems Engineering* **3**(4) 263–279.

Parizi, Mahshid Salemi, Archis Ghate. 2016. Multi-class, multi-resource advance scheduling with no-shows, cancellations and overbooking. *Computers & Operations Research* **67** 90–101.

Pérez, Eduardo, Lewis Ntaimo, César O Malavé, Carla Bailey, Peter McCormack. 2013. Stochastic online appointment scheduling of multi-step sequential procedures in nuclear medicine. *Health Care Management Science* **16**(4) 281–299.

Pinedo, Michael L. 2016. *Scheduling: Theory, Algorithms, and Systems*. Springer, New York.

Popescu, Ioana. 2007. Robust mean-covariance solutions for stochastic optimization. *Operations Research* **55**(1) 98–112.

Porter, Michael E. 2010. What is value in health care? *New England Journal of Medicine* **363**(26) 2477–2481.

Qu, Xiuli, Yidong Peng, Nan Kong, Jing Shi. 2013. A two-phase approach to scheduling multi-category outpatient appointments–a case study of a women's clinic. *Health Care Management Science* **16**(3) 197–216.

Rachuba, Sebastian, Brigitte Werners. 2014. A robust approach for scheduling in hospitals using multiple objectives. *Journal of the Operational Research Society* **65**(4) 546–556.

Roussas, George G. 1997. *A course in mathematical statistics*. Elsevier.

Saghafian, Soroush, Wallace J Hopp, Mark P Van Oyen, Jeffrey S Desmond, Steven L Kronick. 2014. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management* **16**(3) 329–345.

Shu, Jia, Mabel C Chou, Qizhang Liu, Chung-Piaw Teo, I-Lin Wang. 2013. Models for effective deployment and redistribution of bicycles within public bicycle-sharing systems. *Operations Research* **61**(6) 1346–1359.

Steck, GP. 1971. Rectangle probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distribution functions. *The Annals of Mathematical Statistics* **42**(1) 1–11.

Stewart, Moira, Judith Belle Brown, Allan Donner, Ian R. McWhinney, Julian Oates, W. Wayne Weston, John Jordan. 2000. The impact of patient-centered care on outcomes. *Family Practice* **49** 796–804.

Swisher, James R, Sheldon H Jacobson, J Brian Jun, Osman Balci. 2001. Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers & Operations Research* **28**(2) 105–125.

Truong, Van-Anh. 2015. Optimal advance scheduling. *Management Science* **61**(7) 1584–1597.

Wang, Dongyang, Douglas J Morrice, Kumar Muthuraman, Jonathan F Bard, Luci K Leykum, Susan H Noorily. 2018. Coordinated scheduling for a multi-server network in outpatient pre-operative care. *Production and Operations Management* **27**(3) 458–479.

White, Denise L, Craig M Froehle, Kenneth J Klassen. 2011. The effect of integrated scheduling and capacity policies on clinical efficiency. *Production and Operations Management* **20**(3) 442–455.

Wijewickrama, A K A. 2006. Simulation analysis for reducing queues in mixed-patients' outpatient department. *International Journal of Simulation Modelling* **5**(2) 56–68.

Yang, Wenzhuo, Huan Xu. 2016. Distributionally robust chance constraints for non-linear uncertainties. *Mathematical Programming* **155**(1-2) 231–265.

Zhang, Pengfei, Jonathan F Bard, Douglas J Morrice, Karl M Koenig. 2019. Extended open shop scheduling with resource constraints: Appointment scheduling for integrated practice units. *IISE Transactions* **51**(10) 1037–1060.

Vita

Pengfei Zhang was born in Nantong, China. He graduated from the School of the Gifted Young at the University of Science and Technology of China with a Bachelor's degree in Physics. He obtained his Master's degree in Industrial Engineering from the University of Arizona. After that, he worked for the Mayo Clinic in Phoenix as a research assistant intern. He started doctorate studies in Operations Management at the McCombs School of Business at The University of Texas Austin in August, 2015.

When he is not working on research projects and studying, he likes to catch his favourite comedy shows. He completed over ninety 5K runs during his five-year period of stay in Austin.